Approximate Bayesian computation with the Wasserstein distance

Espen Bernton
Department of Statistics, Harvard University, USA.
Pierre E. Jacob
Department of Statistics, Harvard University, USA.
Mathieu Gerber
School of Mathematics, University of Bristol, UK.
Christian P. Robert
CEREMADE, Université Paris-Dauphine and Paris Sciences & Lettres - PSL Research University, France, and Department of Statistics, University of Warwick, UK.

Summary. A growing range of generative statistical models prohibit the numerical evaluation of their likelihood functions. Approximate Bayesian computation (ABC) has become a popular approach to overcome this issue, in which one simulates synthetic data sets given parameters and compares summaries of these data sets with the corresponding observed values. We propose to avoid the use of summaries and the ensuing loss of information by instead using the Wasserstein distance between the empirical distributions of the observed and synthetic data. This generalizes the well-known approach of using order statistics within ABC to arbitrary dimensions. We describe how recently developed approximations of the Wasserstein distance allow the method to scale to realistic data sizes, and propose a new distance based on the Hilbert space-filling curve. We provide a theoretical study of the proposed method, describing consistency as the threshold goes to zero while the observations are kept fixed, and concentration properties as the number of observations grows. Various extensions to time series data are discussed. The approach is illustrated on various examples, including univariate and multivariate g-and-k distributions, a toggle switch model from systems biology, a queueing model, and a Lévy-driven stochastic volatility model.

Keywords: likelihood-free inference, approximate Bayesian computation, Wasserstein distance, optimal transport, generative models

1. Introduction

The likelihood function plays a central role in modern statistics. However, for many models of interest, the likelihood cannot be numerically evaluated. It might still be possible to simulate synthetic data sets from the model given parameters. A popular approach to Bayesian inference in such generative models is approximate Bayesian computation (ABC, Beaumont et al., 2002; Marin et al., 2012). ABC constructs an approximation of the posterior distribution by simulating parameters and synthetic data sets, and retaining the parameters such that the associated data sets are similar enough to the observed data set. Measures of similarity between data sets are often based on summary statistics, such as sample moments. In other words, data sets are considered close if some distance between their summaries is small. The resulting ABC approximations have proven extremely useful, but can lead to a systematic loss of information compared to the original posterior distribution.

We propose here to instead view data sets as empirical distributions and to rely on the Wasserstein distance between synthetic and observed data sets. The Wasserstein distance,

also called the Gini, Mallows, or Kantorovich distance, defines a metric on the space of probability distributions, and has become increasingly popular in statistics and machine learning, due to its appealing computational and statistical properties (e.g. Cuturi, 2013; Srivastava et al., 2015; Sommerfeld and Munk, 2018; Panaretos and Zemel, 2019). We will show that the resulting ABC posterior, which we term the Wasserstein ABC (WABC) posterior, can approximate the posterior distribution arbitrarily well in the limit of the threshold ε going to zero, while bypassing the choice of summaries. Furthermore, we derive asymptotic settings under which the WABC posterior behaves differently from the posterior, illustrating the potential impact of model misspecification and the effect of the dimension of the observation space, by providing upper bounds on concentration rates as the number of observations goes to infinity. The WABC posterior is a particular case of coarsened posterior, and our results are complementary to those of Miller and Dunson (2018).

We further develop two strategies to deal with the specific case of time series. The challenge is that the marginal empirical distributions of time series might not contain enough information to identify all model parameters. In the first approach, which we term curve matching, each data point is augmented with the time at which it was observed. A new ground metric is defined on this extended observation space, which in turn allows for the definition of a Wasserstein distance between time series, with connections to Thorpe et al. (2017). A tuning parameter $\lambda > 0$ allows the proposed distance to approximate the Euclidean distance as $\lambda \to \infty$, and the Wasserstein distance between the marginal distributions as $\lambda \to 0$. The second approach involves transforming the time series such that its empirical distribution contains enough information for parameter estimation. We refer to such transformations as reconstructions and discuss delay reconstructions, as studied in dynamical systems (Stark et al., 2003), and residual reconstructions, as already used in ABC settings (Mengersen et al., 2013).

The calculation of Wasserstein distances is fast for empirical distributions in one dimension, as the main computational task reduces to sorting. For multivariate data sets, we can leverage the rich literature on the computation of Wasserstein distances and approximations thereof (Peyré and Cuturi, 2018). We also propose a new distance utilizing the idea of sorting, termed the Hilbert distance, based on the Hilbert space-filling curve (Sagan, 1994; Gerber and Chopin, 2015). The proposed distance approximates the Wasserstein distance well in low dimensions, but can be computed faster than the exact distance. We also shed light on some theoretical properties of the resulting ABC posterior.

In the following subsections we set up the problem we consider in this work, and briefly introduce ABC and the Wasserstein distance; we refer to Marin et al. (2012) and to Villani (2008) and Santambrogio (2015) for more detailed presentations of ABC and the Wasserstein distance, respectively.

1.1. Set-up, notation and generative models

Throughout this work we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with associated expectation operator $\mathbb{E}$, on which all the random variables are defined. The set of probability measures on a space $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. The data take values in $\mathcal{Y}$, a subset of $\mathbb{R}^{d_y}$ for $d_y \in \mathbb{N}$. We observe $n \in \mathbb{N}$ data points, $y_{1:n} = y_1, \ldots, y_n$, that are distributed according to $\mu_\star^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$. Let $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{y_i}$, where δ_y is the Dirac distribution with mass on $y \in \mathcal{Y}$. With a slight abuse of language, we refer below to $\hat{\mu}_n$ as the empirical distribution of $y_{1:n}$, even in the presence of non i.i.d. observations.

Formally, a model refers to a collection of distributions on $\mathcal{Y}^n$, denoted by $\mathcal{M}^{(n)} = \{\mu_\theta^{(n)} : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y}^n)$, where $\mathcal{H} \subset \mathbb{R}^{d_\theta}$ is the parameter space, endowed with a distance $\rho_\mathcal{H}$ and of dimension $d_\theta \in \mathbb{N}$. However, we will often assume that the sequence of models $(\mathcal{M}^{(n)})_{n \geq 1}$ is such that, for every $\theta \in \mathcal{H}$, the sequence $(\hat{\mu}_{\theta,n})_{n \geq 1}$ of random

probability measures on $\mathcal{Y}$ converges (in some sense) to a distribution $\mu_\theta \in \mathcal{P}(\mathcal{Y})$, where $\hat{\mu}_{\theta,n} = n^{-1}\sum_{i=1}^{n}\delta_{z_i}$ with $z_{1:n} \sim \mu_\theta^{(n)}$. Similarly, we will often assume that $\hat{\mu}_n$ converges to some distribution $\mu_\star \in \mathcal{P}(\mathcal{Y})$ as $n \to \infty$. Whenever the notation $\mu_\star$ and μ_θ is used, it is implicitly assumed that these objects exist. In such cases, we instead refer to $\mathcal{M} = \{\mu_\theta : \theta \in \mathcal{H}\} \subset \mathcal{P}(\mathcal{Y})$ as the model. We say that it is well-specified if there exists $\theta_\star \in \mathcal{H}$ such that $\mu_\star = \mu_{\theta_\star}$; otherwise it is misspecified. Parameters are identifiable if $\theta = \theta'$ is implied by $\mu_\theta = \mu_{\theta'}$.

We consider parameter inference for purely generative models: it is possible to generate observations $z_{1:n}$ from $\mu_\theta^{(n)}$, for all $\theta \in \mathcal{H}$, but it is not possible to numerically evaluate the associated likelihood. In some cases, observations from the model are obtained as $z_{1:n} = g_n(u, \theta)$, where g_n is a known deterministic function and u some known fixed-dimensional random variable independent of θ. Some methods require access to g_n and u (Prangle et al., 2016; Graham and Storkey, 2017); by contrast, here we do not place assumptions on how data sets are generated from the model.

1.2. Approximate Bayesian computation

Let π be a prior distribution on the parameter θ. Consider the following algorithm, where $\varepsilon > 0$ is referred to as the threshold, and $\mathfrak{D}$ denotes a discrepancy measure between two data sets $y_{1:n}$ and $z_{1:n}$, taking non-negative values.

(a) Draw a parameter θ from the prior distribution π, and a synthetic dataset $z_{1:n} \sim \mu_\theta^{(n)}$.
(b) If $\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon$, keep θ, otherwise reject it.

The accepted samples are drawn from the ABC posterior distribution

$$\pi^{\varepsilon}_{y_{1:n}}(d\theta) = \frac{\pi(d\theta)\int_{\mathcal{Y}^n}\mathbb{1}\left(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\right)\mu_\theta^{(n)}(dz_{1:n})}{\int_{\mathcal{H}}\pi(d\theta)\int_{\mathcal{Y}^n}\mathbb{1}\left(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\right)\mu_\theta^{(n)}(dz_{1:n})}, \tag{1}$$

where $\mathbb{1}$ is the indicator function. A more sophisticated algorithm to approximate ABC posteriors, which we will apply in our numerical experiments, is described in Section 2.1.

Let ρ be a distance on the observation space $\mathcal{Y}$, referred to as the ground distance. Suppose that $\mathfrak{D}$ is chosen as

$$\mathfrak{D}(y_{1:n}, z_{1:n})^p = \frac{1}{n}\sum_{i=1}^{n}\rho(y_i, z_i)^p. \tag{2}$$

Then, the resulting ABC posterior can be shown to have the desirable theoretical property of converging to the standard posterior as $\varepsilon \to 0$ (Prangle et al., 2016, see also Proposition 3.1). In the case where $p = 2$, $\mathcal{Y} \subset \mathbb{R}$, and $\rho(y_i, z_i) = |y_i - z_i|$, $\mathfrak{D}$ is a scaled version of the Euclidean distance between the vectors $y_{1:n}$ and $z_{1:n}$.

However, this approach is in most cases impractical due to the large variation of $\mathfrak{D}(y_{1:n}, z_{1:n})$ over repeated samples from $\mu_\theta^{(n)}$. A rare example of practical use of ABC with the Euclidean distance is given in Sousa et al. (2009). A large proportion of the ABC literature is devoted to studying ABC posteriors in the setting where $\mathfrak{D}$ is the Euclidean distance between summaries, i.e. $\mathfrak{D}(y_{1:n}, z_{1:n}) = \|\eta(y_{1:n}) - \eta(z_{1:n})\|$, where $\eta : \mathcal{Y}^n \to \mathbb{R}^{d_\eta}$ for some small d_η. Using summaries can lead to a loss of information: the resulting ABC posterior converges, at best, to the conditional distribution of θ given $\eta(y_{1:n})$, as $\varepsilon \to 0$. A trade-off ensues, where using more summaries reduces the information loss, but increases the variation in the distance over repeated model simulations (Fearnhead and Prangle, 2012).

1.3. Wasserstein distance

A natural approach to reducing the variance of the distance defined in (2), while hoping to avoid the loss of information incurred by the use of summary statistics, is to instead consider the distance

$$\mathfrak{W}_p(y_{1:n}, z_{1:n})^p = \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, z_{\sigma(i)})^p, \tag{3}$$

where $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$. Indeed, when the observations are univariate and $\rho(y_i, z_j) = |y_i - z_j|$, the above infimum is achieved by sorting $y_{1:n}$ and $z_{1:n}$ in increasing order and matching the order statistics. Using order statistics as a choice of summary within ABC has been suggested multiple times in the literature, see e.g. Sousa et al. (2009); Fearnhead and Prangle (2012). It turns out that $\mathfrak{W}_p(y_{1:n}, z_{1:n})$ is the p-Wasserstein distance between the empirical distributions supported on the data sets $y_{1:n}$ and $z_{1:n}$. From this perspective, our proposal of using the Wasserstein distance between empirical distributions can be thought of as generalizing the use of order statistics within ABC to arbitrary dimensions.

More formally, let $\mathcal{P}_p(\mathcal{Y})$ with $p \geq 1$ (e.g. $p = 1$ or 2) be the set of distributions $\mu \in \mathcal{P}(\mathcal{Y})$ with finite p-th moment: there exists $y_0 \in \mathcal{Y}$ such that $\int_{\mathcal{Y}} \rho(y, y_0)^p d\mu(y) < \infty$. The space $\mathcal{P}_p(\mathcal{Y})$ is referred to as the p-Wasserstein space of distributions on $\mathcal{Y}$ (Villani, 2008). The p-Wasserstein distance is a finite metric on $\mathcal{P}_p(\mathcal{Y})$, defined by the transport problem

$$\mathfrak{W}_p(\mu, \nu)^p = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} \rho(x, y)^p d\gamma(x, y), \tag{4}$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathcal{Y} \times \mathcal{Y}$ with marginals μ and ν respectively; see the notes in Chapter 6 of Villani (2008) for a brief history of this distance and its central role in optimal transport.

As in (3), we also write $\mathfrak{W}_p(y_{1:n}, z_{1:m})$ for $\mathfrak{W}_p(\hat{\mu}_n, \hat{\nu}_m)$, where $\hat{\mu}_n$ and $\hat{\nu}_m$ stand for the empirical distributions $n^{-1}\sum_{i=1}^{n} \delta_{y_i}$ and $m^{-1}\sum_{i=1}^{m} \delta_{z_i}$. In particular, the Wasserstein distance between two empirical distributions with unweighted atoms takes the form

$$\mathfrak{W}_p(y_{1:n}, z_{1:m})^p = \inf_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho(y_i, z_j)^p \gamma_{ij} \tag{5}$$

where $\Gamma_{n,m}$ is the set of $n \times m$ matrices with non-negative entries, columns summing to m^{-1}, and rows summing to n^{-1}. We focus on the case $n = m$, for which it is known that the solution to the optimization problem, $\gamma^\star$, corresponds to an assignment matrix with only one non-zero entry per row and column, equal to n^{-1} (see e.g. the introductory chapter in Villani, 2003). In this special case, the Wasserstein distance can thus be represented as in (3). Computing the Wasserstein distance between two samples of the same size can therefore also be thought of as a matching problem; see Section 2.3.

1.4. Related works and plan

The minimum Wasserstein estimator (MWE), first studied in Bassetti et al. (2006), is an example of a minimum distance estimator (Basu et al., 2011) and is defined as $\hat{\theta}_n = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$. To extend this approach to generative models, Bernton et al. (2017) introduce the minimum expected Wasserstein estimator (MEWE), defined as $\hat{\theta}_{n,m} = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathbb{E}\left[\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})\right]$, where the expectation refers to the distribution of $z_{1:m} \sim \mu_\theta^{(m)}$. General results on both the MWE and MEWE are obtained in the technical report of Bernton et al. (2017). Another method related to our approach was proposed by Park et al. (2016), who bypass the choice of summary statistics in the definition of

the ABC posterior in (1) by using a discrepancy measure $\mathfrak{D}$ such that $\mathfrak{D}(y_{1:n}, z_{1:n})$ is an estimate of the maximum mean discrepancy (MMD) between $\hat{\mu}_n$ and $\mu_\theta^{(n)}$.

Our contributions are structured as follows: the proposed approach to Bayesian inference in generative models using the Wasserstein distance is described in Section 2, some theoretical properties of the Wasserstein ABC posterior is detailed in Section 3, methods to handle time series are proposed in Section 4, and numerical illustrations in Section 5, where in each example we make comparisons to existing methods, such as semi-automatic ABC (Fearnhead and Prangle, 2012). The code is available on GitHub at github.com/pierrejacob/winference. The supplementary material includes additional theoretical results and details on computational aspects, as referenced in the present article.

2. Wasserstein ABC

The distribution $\pi^{\varepsilon}_{y_{1:n}}(d\theta)$ of (1), with $\mathfrak{D}$ replaced by $\mathfrak{W}_p$ for some choice of $p \geq 1$, is referred to as the Wasserstein ABC (WABC) posterior; that is, the WABC posterior is defined by

$$\pi^{\varepsilon}_{y_{1:n}}(d\theta) = \frac{\pi(d\theta) \int_{\mathcal{Y}^n} \mathbb{1}\left(\mathfrak{W}_p(y_{1:n}, z_{1:n}) \leq \varepsilon\right) \mu_\theta^{(n)}(dz_{1:n})}{\int_{\mathcal{H}} \pi(d\theta) \int_{\mathcal{Y}^n} \mathbb{1}\left(\mathfrak{W}_p(y_{1:n}, z_{1:n}) \leq \varepsilon\right) \mu_\theta^{(n)}(dz_{1:n})} \tag{6}$$

with $\mathfrak{W}_p(y_{1:n}, z_{1:n})$ defined in (3). Throughout the experiments of this article we set $p = 1$, which makes minimal assumptions on the existence of moments of the data-generating process.

As mentioned in the introductory section, the motivation for choosing $\mathfrak{D}$ to be the Wasserstein distance is to have a discrepancy measure $\mathfrak{D}(y_{1:n}, z_{1:n})$ that has both a small variance and results in an ABC posterior that has satisfactory theoretical properties. In particular, we show in Section 3 that, as per ABC based on the Euclidean distance, the WABC posterior converges to the true posterior distribution as $\varepsilon \to 0$. In that section, we also provide a result showing that, as $n \to \infty$ and the threshold ε converges slowly enough to some minimal value $\varepsilon_\star \geq 0$, the WABC posterior concentrates around $\theta_\star := \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$. In the well-specified case, $\theta_\star$ coincides with the data-generating parameter. In the misspecified case, $\theta_\star$ is typically different from where the actual posterior concentrates, which is around the minimizer of $\theta \mapsto \mathrm{KL}(\mu_\star | \mu_\theta)$, where KL refers to the Kullback–Leibler divergence. The experiments in Section 5 contains examples where the WABC posterior provides a practical and accurate approximation of the standard posterior, and examples where it does not, partly because of the computational difficulty of sampling from the WABC posterior when ε is small.

2.1. Sampling sequentially from the WABC posterior

Instead of the rejection sampler of Section 1.2, we will target the WABC and other ABC posteriors using a sequential Monte Carlo (SMC) approach, with N particles exploring the parameter space (Del Moral et al., 2012). The algorithm starts with a threshold $\varepsilon_0 = +\infty$, for which the WABC posterior is the prior. Given the Monte Carlo approximation of the WABC posterior for ε_{t-1}, the next value ε_t is chosen so as to maintain a number of unique particles of at least αN, with $\alpha \in (0, 1]$. Upon choosing ε_t, resampling and rejuvenation steps are triggered and the algorithm proceeds. In the experiments, we will run the algorithm until a fixed budget of model simulations is reached. At the end of the run, the algorithm provides N parameter samples and synthetic data sets, associated with a threshold ε_T.

The algorithm is parallelizable over the N particles, and thus over equally many model simulations and distance calculations. Any choice of MCMC kernel can be used within the rejuvenation steps. In particular, we use the r-hit kernel of Lee (2012), shown to be

advantageous compared to standard ABC-MCMC kernels in Lee and Łatuszyński (2014). We choose the number of hits to be 2 by default. For the proposals of the MCMC steps, we use a mixture of multivariate Normal distributions, with 5 components by default. We set N to be $2,048$ and α to be 50%. These default tuning parameters are used throughout all the numerical experiments of Section 5, unless otherwise specified. Full details on the SMC algorithm are given in the supplementary material.

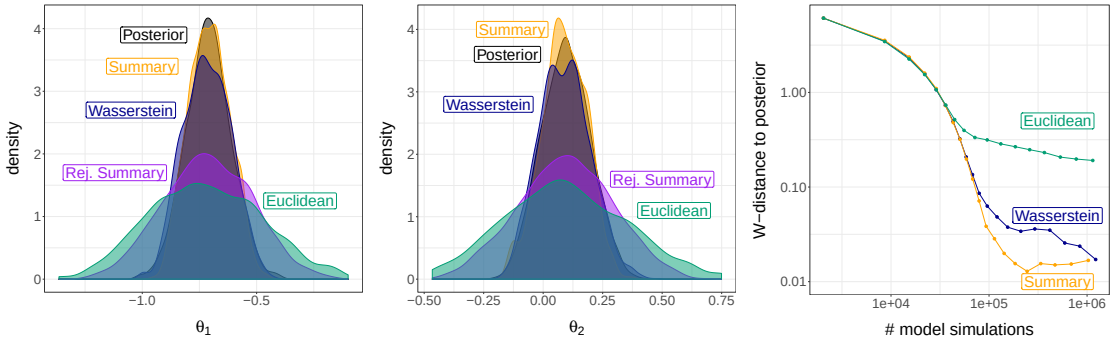
2.2. Illustration on a Normal location model

Consider 100 i.i.d. observations generated from a bivariate Normal distribution. The mean components are drawn from a standard Normal distribution, and the generated values are approximately -0.71 and 0.09. The covariance is equal to 1 on the diagonal and 0.5 off the diagonal. The parameter θ is the mean vector, and is assigned a centered Normal prior with variance 25 on each component.

We compare WABC with two other methods: ABC using the Euclidean distance between the data sets, and ABC using the Euclidean distance between sample means, which for this model are sufficient summary statistics. All three ABC posteriors are approximated using the SMC sampler described in Section 2.1. The summary-based ABC posterior is also approximated using the simple rejection sampler given in Section 1.2 to illustrate the benefit of the SMC approach. All methods are run for a budget of 10^6 model simulations, using $N = 2,048$ particles in the SMC sampler. The rejection sampler accepted only the $2,048$ draws yielding the smallest distances. Approximations of the marginal posterior distributions of the parameters are given in Figures 1a and 1b, illustrating that the SMC-based ABC methods with the Wasserstein distance and with sufficient statistics both approximate the posterior accurately.

To quantify the difference between the obtained ABC samples and the posterior, we again use the Wasserstein distance. Specifically, we independently draw $2,048$ samples from the posterior distribution, and compute the Wasserstein distance between these samples and the $N = 2,048$ ABC samples produced by the SMC algorithm. We plot the resulting distances against the number of model simulations in Figure 1c, in log-log scale. As expected, ABC with sufficient statistics converges fastest to the posterior. It should be noted that sufficient statistics are almost never available in realistic applications of ABC. The proposed WABC approach performs almost as well, but requires more model simulations to yield comparable results. In contrast, the ABC approach with the Euclidean distance struggles to approximate the posterior accurately. Extrapolating from the plot, it would seemingly take billions of model simulations for the latter ABC approach to approximate the posterior as accurately as the other two methods. Similarly, despite being based on the sufficient statistic, the rejection sampler does not adequately estimate the posterior distribution for the given sample budged. The estimated 1-Wasserstein distance between the $2,048$ accepted samples and the posterior was 0.63.

In terms of computing time, based on our R implementation on an Intel Core i5 (2.5GHz), simulating a data set took on average $4.0 \times 10^{-5} s$. Computing the discrepancy between data sets took on average $6.4 \times 10^{-5} s$ for the summary-based distance, $3.8 \times 10^{-4} s$ for the Euclidean distance, and $1.2 \times 10^{-2} s$ for the Wasserstein distance; see Section 2.3 for fast approximations of the Wasserstein distance. The SMC sampler is algorithmically more involved than the rejection sampler, and one could ask whether the added computational effort is justified. In this example, the total time required by the SMC algorithm using the summary statistic was $169s$, whereas the analogous rejection sampler took $141s$. This illustrates that even when one can very cheaply simulate data and compute distances, the added costs associated with an SMC sampler are relatively small; see Del Moral et al. (2012); Filippi et al. (2013); Sisson et al. (2018) for more details on SMC samplers for ABC purposes.



(a) Posteriors of θ_1.

(b) Posteriors of θ_2.

(c) $\mathfrak{W}_1$-distance to posterior, versus number of model simulations (log-log scale).

Fig. 1: ABC in the bivariate Normal location model of Section 2.2. ABC approximations of the posterior after 10^6 model simulations (left and middle), overlaid the actual posterior. On the right, the Wasserstein distance between ABC posterior samples and exact posterior samples is plotted against the number of model simulations (in log-log scale). In principle, these ABC approximations converge to the posterior as $\varepsilon \to 0$. Yet, for a given number of model simulations, the quality of the ABC approximation is sensitive to the choice of distance and sampling algorithm.

2.3. *Computing and approximating the Wasserstein distance*

Computing the Wasserstein distance between the distributions $\hat{\mu}_n = n^{-1}\sum_{i=1}^n \delta_{y_i}$ and $\hat{\nu}_n = n^{-1}\sum_{i=1}^n \delta_{z_i}$ reduces to a linear sum assignment problem, as in (3). In the univariate case, finding the optimal permutation can be done by sorting the vectors $y_{1:n}$ and $z_{1:n}$ in increasing order, obtaining the orders $\sigma_y(i)$ and $\sigma_z(i)$ for $i \in \{1, \ldots, n\}$. Then, one associates each y_i with $z_{\sigma(i)}$ where $\sigma(i) = \sigma_z \circ \sigma_y^{-1}(i)$. The cost of the Wasserstein distance computation is thus of order $n \log n$ for distributions on one-dimensional spaces.

In multivariate settings, (3) can be solved by the Hungarian algorithm for a cost of order n^3. Other algorithms have a cost of order $n^{2.5}\log(n\,C_n)$, with $C_n = \max_{1\leq i,j\leq n}\rho(y_i, z_j)$, and can therefore be more efficient when C_n is small (Burkard et al., 2009, Section 4.1.3). In our numerical experiments, we use the short-list method presented in Gottschlich and Schuhmacher (2014) and implemented in Schuhmacher et al. (2017). This simplex algorithm-derived method comes without guarantees of polynomial running times, but Gottschlich and Schuhmacher (2014) show empirically that their method tends to have sub-cubic cost.

The cubic cost of computing Wasserstein distances in the multivariate setting can be prohibitive for large data sets. However, many applications of ABC involve relatively small numbers of observations from complex models which are expensive to simulate. In these settings, the cost of simulating synthetic data sets might dominate the model-free cost of computing distances. Note also that the dimension d_y of the observation space only enters the ground distance ρ, and thus the cost of computing the Wasserstein distance under a Euclidean ground metric is linear in d_y.

2.3.1. *Fast approximations*

In conjunction with its increasing popularity as a tool for inference in statistics and machine learning, there has been a rapid growth in the number of algorithms that approximate the Wasserstein distance at reduced computational costs; see Peyré and Cuturi (2018). In particular, they provide an in-depth discussion of the popular method proposed by Cuturi (2013), in which the optimization problem in (5) is regularized using an entropic constraint on the joint distribution γ. Consider $\gamma^\zeta = \operatorname{argmin}_{\gamma\in\Gamma_n}\sum_{i,j=1}^n \rho(y_i, z_j)^p\gamma_{ij} +$

$\zeta \sum_{i,j=1}^n \gamma_{ij} \log \gamma_{ij}$, which includes a negative penalty on the entropy of γ, and define the dual-Sinkhorn divergence $S_p^\zeta(y_{1:n}, z_{1:n})^p = \sum_{i,j=1}^n \rho(y_i, z_j)^p \gamma_{ij}^\zeta$. The regularized problem can be solved iteratively by Sinkhorn's algorithm, which involves matrix-vector multiplications resulting in a cost of order n^2 per iteration. If $\zeta \to 0$, the dual-Sinkhorn divergence converges to the Wasserstein distance, whereas if $\zeta \to \infty$ it converges to the maximum mean discrepancy (Ramdas et al., 2017). It can therefore be seen as an interpolation between optimal transport and kernel-based distances. Further properties of the dual-Sinkhorn divergence and other algorithms to approximate it are discussed in Peyré and Cuturi (2018).

Unlike the optimal coupling that yields the exact Wasserstein distance, the coupling obtained in the regularized problem is typically not an assignment matrix. In the following subsections, we discuss two simple approaches with different computational complexities that yield couplings that are assignments. This has the benefit of aiding the theoretical analysis in Section 3.

2.3.2. Hilbert distance

The assignment problem in (3) can be solved in $n \log n$ in the univariate case by sorting the samples. We propose a new distance generalizing this idea when $d_y > 1$, by sorting samples according to their projection via the Hilbert space-filling curve. As shown in Gerber and Chopin (2015) and Schretter et al. (2016), transformations through the Hilbert space-filling curve and its inverse preserve a notion of distance between probability measures. The Hilbert curve $H : [0,1] \to [0,1]^{d_y}$ is a Hölder continuous mapping from $[0,1]$ into $[0,1]^{d_y}$. One can define a measurable pseudo-inverse $h : [0,1]^{d_y} \to [0,1]$ verifying $h(H(x)) = x$ for all $x \in [0,1]$ (Gerber et al., 2019). We assume in this subsection that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ is such that there exists a mapping $\psi : \mathcal{Y} \to (0,1)^{d_y}$ verifying, for $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$, $\psi(y) = \big(\psi_1(y_1), ..., \psi_{d_y}(y_{d_y})\big)$ where the ψ_i's are continuous and strictly monotone. For instance, if $\mathcal{Y} = \mathbb{R}^{d_y}$, one can take ψ to be the component-wise logistic transformation; see Gerber and Chopin (2015) for more details. By construction, the mapping $h_\mathcal{Y} := h \circ \psi : \mathcal{Y} \to (0,1)$ is one-to-one. For two vectors $y_{1:n}$ and $z_{1:n}$, denote by σ_y and σ_z the permutations obtained by mapping the vectors through $h_\mathcal{Y}$ and sorting the resulting univariate vectors in increasing order. We define the Hilbert distance $\mathfrak{H}_p$ between the empirical distributions of $y_{1:n}$ and $z_{1:n}$ by

$$\mathfrak{H}_p(y_{1:n}, z_{1:n})^p = \frac{1}{n} \sum_{i=1}^n \rho\big(y_i, z_{\sigma(i)}\big)^p, \tag{7}$$

where $\sigma(i) = \sigma_z \circ \sigma_y^{-1}(i)$ for all $i \in \{1, \ldots, n\}$.

PROPOSITION 2.1. *For any integer $n \geq 1$ and real number $p \geq 1$, $\mathfrak{H}_p$ defines a distance on the space of empirical distributions of size n.*

The Hilbert distance can be computed at a cost in the order of $n \log n$ and an implementation is provided by the function `hilbert_sort` in The Computational Geometry Algorithms Library (2016). From a practical point of view, this implementation has the attractive property of not having to map the samples to $(0,1)^{d_y}$ and hence having to choose a specific mapping ψ. Instead, this function directly constructs the Hilbert curve around the input point set.

Despite not being defined in terms of a transport problem, the Hilbert distance yields approximations of the Wasserstein distance that are accurate for small d_y, as illustrated in the supplementary material. More importantly for its use within ABC, the level sets of the (random) map $\theta \mapsto \mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n})$ appear to be close to those of the analogous Wasserstein

distance. The two distances therefore discriminate between parameters in similar fashions. However, this behavior tends to deteriorate as the dimension d_y grows.

The coupling produced by Hilbert sorting is feasible for the assignment problem in (3). Therefore, it is always greater than the Wasserstein distance, which minimizes the objective therein. This property plays an important role in showing that the ABC posterior based on the Hilbert distance concentrates on $\theta_\star$ as $n \to \infty$ and the threshold ε decreases sufficiently slowly. In the supplementary materials, we provide such a result under the assumption that the model is well-specified, but leave further theoretical analysis under milder conditions for future research. Other one-dimensional projections of multivariate samples, followed by Wasserstein distance computation using the projected samples, have been proposed in the computational optimal transport literature (Rabin et al., 2011; Bonneel et al., 2015), also leading to computational costs in $n \log n$.

2.3.3. *Swapping distance*

Viewing the Wasserstein distance calculation as the assignment problem in (3), Puccetti (2017) proposed a greedy swapping algorithm to approximate the optimal assignment. Consider an arbitrary permutation σ of $\{1, \ldots, n\}$, and the associated transport cost $\sum_{i=1}^n \rho(y_i, z_{\sigma(i)})^p$. The swapping algorithm consists in checking, for all $1 \leq i < j \leq n$, whether $\rho(y_i, z_{\sigma(i)})^p + \rho(y_j, z_{\sigma(j)})^p$ is less or greater than $\rho(y_i, z_{\sigma(j)})^p + \rho(y_j, z_{\sigma(i)})^p$. If it is greater, then one swaps $\sigma(i)$ and $\sigma(j)$, resulting in a decrease of the transport cost. One can repeat these sweeps over $1 \leq i < j \leq n$, until the assignment is left unchanged, and denote it by $\tilde{\sigma}$. Each sweep has a cost of order n^2 operations. There is no guarantee that the resulting assignment $\tilde{\sigma}$ corresponds to the optimal one. Note that we initialize the algorithm with the assignment obtained by Hilbert sorting for a negligible cost of $n \log n$. We refer to the resulting distance $(n^{-1} \sum_{i=1}^n \rho(y_i, z_{\tilde{\sigma}(i)})^p)^{1/p}$ as the swapping distance.

The swapping distance between $y_{1:n}$ and $z_{1:n}$ takes values that are, by construction, between the Wasserstein distance $\mathfrak{W}_p(y_{1:n}, z_{1:n})$ and the Hilbert distance $\mathfrak{H}_p(y_{1:n}, z_{1:n})$. Thanks to this property, we show in the supplementary material that the associated ABC posterior concentrates on $\theta_\star$ as $n \to \infty$ and the threshold ε decreases sufficiently slowly. As with the Hilbert distance, this result is obtained under the assumption that the model is well-specified and leave further theoretical analysis under milder conditions for future research. In the supplementary material, we also observe that the swapping distance can approximate the Wasserstein distance more accurately than the Hilbert distance as the dimension d_y grows.

2.3.4. *Sub-sampling*

Any of the aforementioned distances can be computed faster by first sub-sampling $m < n$ points from $y_{1:n}$ and $z_{1:n}$, and then computing the distance between the resulting distributions. This increases the variance of the calculated distances, introducing a trade-off with computation time. In the case of the Wasserstein distance, this approach could be studied formally using the results of Sommerfeld and Munk (2018). Other multiscale approaches can also be used to accelerate computation (Mérigot, 2011). We remark that computing the distance between vectors containing subsets of order statistics (Fearnhead and Prangle, 2012) can be viewed as an example of a multiscale approach to approximating the Wasserstein distance.

2.3.5. *Combining distances*

It might be useful to combine distances. For instance, one might want to start exploring the parameter space with a cheap approximation, and switch to the exact Wasserstein

distance in a region of interest; or use the cheap approximation to save computations in a delayed acceptance scheme. One might also combine a transport distance with a distance between summaries. We can combine distances in the ABC framework by introducing a threshold for each distance, and define the ABC posterior as in (1), with a product of indicators corresponding to each distance. We explore the combination of distances in the numerical experiments of Section 5.4.

3. Theoretical properties

We study the behavior of the Wasserstein ABC posterior under different asymptotic regimes. First, we give conditions on a discrepancy measure for the associated ABC posterior to converge to the posterior as the threshold ε goes to zero, while keeping the observed data fixed. We then discuss the behavior of the WABC posterior as $n \to \infty$ for fixed $\varepsilon > 0$. Finally, we establish bounds on the rates of concentration of the WABC posterior as the data size n grows and the threshold ε shrinks sufficiently slowly at a rate dependent on n, similar to Frazier et al. (2018) in the case of summary-based ABC. Proofs are deferred to the appendix.

We remark that the assumptions underlying our results are typically hard to check in practice, due to the complexity and intractable likelihoods of the models to which ABC methods are applied. This is also true for state-of-the-art asymptotic results about summary-based ABC methods, which, for example, require injectivity and growth conditions on the "binding function" to which the summary statistics converge (Frazier et al., 2018). Nonetheless, we believe that our results provide insight into the statistical properties of the WABC posterior. For instance, in Corollary 3.1 we give conditions under which the WABC posterior concentrates around $\theta_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\theta, \mu_\star)$ as n grows. When the model is misspecified, this is in contrast with the posterior, which is known to concentrate around $\operatorname{argmin}_{\theta \in \mathcal{H}} \mathrm{KL}(\mu_\theta, \mu_\star)$ (see e.g. Müller, 2013).

3.1. Behavior as $\varepsilon \to 0$ for fixed observations

The following result establishes conditions under which a non-negative measure of discrepancy between data sets $\mathfrak{D}$ yields an ABC posterior that converges to the true posterior as $\varepsilon \to 0$, while the observations are kept fixed.

PROPOSITION 3.1. *Suppose that $\mu_\theta^{(n)}$ has a continuous density $f_\theta^{(n)}$ and that*

$$\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_\mathcal{H}} f_\theta^{(n)}(y_{1:n}) < \infty,$$

where $\mathcal{N}_\mathcal{H}$ is a set such that $\pi(\mathcal{N}_\mathcal{H}) = 0$. Suppose that there exists $\bar{\varepsilon} > 0$ such that

$$\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_\mathcal{H}} \sup_{z_{1:n} \in \mathcal{A}^{\bar{\varepsilon}}} f_\theta^{(n)}(z_{1:n}) < \infty,$$

where $\mathcal{A}^{\bar{\varepsilon}} = \{z_{1:n} : \mathfrak{D}(y_{1:n}, z_{1:n}) \leq \bar{\varepsilon}\}$. Suppose also that $\mathfrak{D}$ is continuous in the sense that $\mathfrak{D}(y_{1:n}, z_{1:n}) \to \mathfrak{D}(y_{1:n}, x_{1:n})$ whenever $z_{1:n} \to x_{1:n}$ component-wise in the metric ρ. If either

(a) $f_\theta^{(n)}$ is n-exchangeable, such that $f_\theta^{(n)}(y_{1:n}) = f_\theta^{(n)}(y_{\sigma(1:n)})$ for any $\sigma \in \mathcal{S}_n$, and $\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, or
(b) $\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{1:n}$,

then, keeping $y_{1:n}$ fixed, the ABC posterior converges strongly to the posterior as $\varepsilon \to 0$.

The Wasserstein distance applied to unmodified data satisfies $\mathfrak{W}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, making condition (a) of Proposition 3.1 applicable. In Section 4, we will discuss two methods applicable to time series that lead to discrepancies for which condition (b) holds. Note that this result does not guarantee that the Monte Carlo algorithm employed to sample the ABC posterior distribution, with an adaptive mechanism to decrease the threshold, will be successful at reaching low thresholds in a reasonable time.

3.2. Behavior as $n \to \infty$ for fixed ε

Under weak conditions, the WABC posterior distribution $\pi^{\varepsilon}_{y_{1:n}}(d\theta)$ in (1) converges to $\pi(d\theta|\mathfrak{W}_p(\mu_\theta, \mu_\star) < \varepsilon)$ as $n \to \infty$ for a fixed threshold ε, following the reasoning in Miller and Dunson (2018) for general weakly-continuous distances, which include the Wasserstein distance. Therefore, the WABC distribution with a fixed ε does not converge to a Dirac mass, contrarily to the standard posterior. As argued in Miller and Dunson (2018), this can have some benefit in case of model misspecification: the WABC posterior is less sensitive to perturbations of the data-generating process than the standard posterior.

3.3. Concentration as n increases and ε decreases

A sequence of distributions $\pi_{y_{1:n}}$ on $\mathcal{H}$, depending on the data $y_{1:n}$, is consistent at $\theta_\star$ if, for any $\delta > 0$, $\mathbb{E}[\pi_{y_{1:n}}(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_\star) > \delta\})] \to 0$, where the expectation is taken with respect to $\mu_\star^{(n)}$. Finding rates of concentration for $\pi_{y_{1:n}}$ involves finding the fastest decaying sequence $\delta_n > 0$ such that the limit above holds. More precisely, we say that the rate of concentration of $\pi_{y_{1:n}}$ is bounded above by the sequence δ_n if $\mathbb{E}[\pi_{y_{1:n}}(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_\star) > \delta_n\})] \to 0$.

We establish upper bounds on the rates of concentration of the sequence of WABC posteriors around $\theta_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\theta, \mu_\star)$, as the data size n grows and the threshold shrinks slowly towards $\varepsilon_\star = \mathfrak{W}_p(\mu_{\theta_\star}, \mu_\star)$ at a rate dependent on n. Although we focus on the Wasserstein distance in this section, the reasoning also holds for other metrics on $\mathcal{P}(\mathcal{Y})$; see Section 2.3 and the supplementary material.

Our first assumption is on the convergence of the empirical distribution of the data.

ASSUMPTION 1. *The data-generating process is such that* $\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \to 0$, *in* $\mathbb{P}$-*probability, as* $n \to \infty$.

In the supplementary material, we derive a few different conditions under which Assumption 1 holds for i.i.d. data and certain classes of dependent processes. Additionally, the moment and concentration inequalities of Fournier and Guillin (2015); Weed and Bach (2017) can also be used to verify both this and the next assumption.

ASSUMPTION 2. *For any* $\varepsilon > 0$, $\mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \varepsilon) \leq c(\theta) f_n(\varepsilon)$, *where* $f_n(\varepsilon)$ *is a sequence of functions that are strictly decreasing in* ε *for fixed* n *and* $f_n(\varepsilon) \to 0$ *for fixed* ε *as* $n \to \infty$. *The function* $c : \mathcal{H} \to \mathbb{R}^+$ *is* π-*integrable, and satisfies* $c(\theta) \leq c_0$ *for some* $c_0 > 0$, *for all* θ *such that, for some* $\delta_0 > 0$, $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \delta_0 + \varepsilon_\star$.

For well-specified models, note that Assumption 2 implies Assumption 1. The next assumption states that the prior distribution puts enough mass on the sets of parameters θ that yield distributions μ_θ close to $\mu_\star$ in the Wasserstein distance.

ASSUMPTION 3. *There exist* $L > 0$ *and* $c_\pi > 0$ *such that, for all* ε *small enough,*

$$\pi(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon + \varepsilon_\star\}) \geq c_\pi \varepsilon^L.$$

The main result of this subsection is on the concentration of the WABC posteriors on the aforementioned sets.

PROPOSITION 3.2. *Under Assumptions 1-3, consider a sequence* $(\varepsilon_n)_{n\geq 0}$ *such that, as* $n \to \infty$, $\varepsilon_n \to 0$, $f_n(\varepsilon_n) \to 0$, *and* $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon_n) \to 1$. *Then, the WABC posterior with threshold* $\varepsilon_n + \varepsilon_\star$ *satisfies, for some* $0 < C < \infty$ *and any* $0 < R < \infty$,

$$\pi_{y_{1:n}}^{\varepsilon_n+\varepsilon_\star}\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > \varepsilon_\star + 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R)\}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$*-probability going to* 1 *as* $n \to \infty$.

The assumptions that $f_n(\varepsilon_n) \to 0$ and that $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon_n) \to 1$ imply that ε_n has to be the slowest of the two convergence rates: that of $\hat{\mu}_n$ to $\mu_\star$ and that of $\hat{\mu}_{\theta,n}$ to μ_θ. We can further relate concentration on the sets $\{\theta : \mathfrak{W}_p(\mu_\theta, \mu_\star) < \delta' + \varepsilon_\star\}$, for some $\delta' > 0$, to concentration on the sets $\{\theta : \rho_{\mathcal{H}}(\theta, \theta_\star) < \delta\}$, for some $\delta > 0$, assuming the parameter $\theta_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is well-defined. In turn, this leads to concentration rates of the WABC posteriors. To that end, consider the following assumptions.

ASSUMPTION 4. *The parameter* $\theta_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ *exists, and is well-separated in the sense that, for all* $\delta > 0$, *there exists* $\delta' > 0$ *such that*

$$\inf_{\{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)>\delta\}} \mathfrak{W}_p(\mu_\theta, \mu_\star) > \mathfrak{W}_p(\mu_{\theta_\star}, \mu_\star) + \delta'.$$

This assumption is akin to those made in the study of the asymptotic properties of the maximum likelihood estimator under misspecification, where $\theta_\star$ is defined in terms of the Kullback–Leibler divergence. In the supplementary material, we give a proposition establishing conditions under which Assumption 4 holds.

Under Assumption 4, note that the last part of Assumption 2 is implied by $c(\theta) \leq c_0$ for all θ with $\mathfrak{W}_p(\mu_{\theta_\star}, \mu_\theta) \leq \delta_0$, for some $\delta_0 > 0$. Indeed, $\mathfrak{W}_p(\mu_{\theta_\star}, \mu_\theta) \leq \delta_0$ implies that $\mathfrak{W}_p(\mu_\theta, \mu_\star) - \mathfrak{W}_p(\mu_\star, \mu_{\theta_\star}) \leq \delta_0$. Since $\varepsilon_\star = \mathfrak{W}_p(\mu_\star, \mu_{\theta_\star})$, the argument follows. By the same reasoning, Assumption 3 is implied by $\pi(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\}) \geq c_\pi \varepsilon^L$, for some $c_\pi > 0$ and $L > 0$.

ASSUMPTION 5. *The parameters are identifiable, and there exist* $K > 0$, $\alpha > 0$ *and an open neighborhood* $U \subset \mathcal{H}$ *of* $\theta_\star$, *such that, for all* $\theta \in U$,

$$\rho_{\mathcal{H}}(\theta, \theta_\star) \leq K(\mathfrak{W}_p(\mu_\theta, \mu_\star) - \varepsilon_\star)^\alpha.$$

COROLLARY 3.1. *Under Assumptions 1-5, consider a sequence* $(\varepsilon_n)_{n\geq 0}$ *such that, as* $n \to \infty$, $\varepsilon_n \to 0$, $f_n(\varepsilon_n) \to 0$, $f_n^{-1}(\varepsilon_n^L) \to 0$ *and* $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon_n) \to 1$. *Then the WABC posterior with threshold* $\varepsilon_n + \varepsilon_\star$ *satisfies, for some* $0 < C < \infty$ *and any* $0 < R < \infty$,

$$\pi_{y_{1:n}}^{\varepsilon_n+\varepsilon_\star}\left(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta, \theta_\star) > K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^\alpha\}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$*-probability going to* 1.

This result bounds the concentration rate from above through the expression $\delta_n = K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^\alpha$, but we remark that it is not clear whether this bound is optimal in any sense. Explicit upper bounds for certain classes of models and data-generating processes, such as location-scale models and the AR(1) model in Example 4.2, are given in the supplementary material. Important aspects of the method that appear in these bounds include the dimension of the observation space $\mathcal{Y}$, the order p of the Wasserstein distance, and model misspecification, through the exponent α in Assumption 5.

The result provides some insight into the behavior of the method when ε_n converges *slowly* to $\varepsilon_\star$. However, it is unclear what happens when ε_n decays to a value smaller than $\varepsilon_\star$ at a rate faster than that prescribed by Corollary 3.1. As shown in Proposition 3.1, the WABC posterior converges to the true posterior when $\varepsilon \to 0$ for fixed observations. The posterior itself is known to concentrate around the point in $\mathcal{H}$ minimizing the KL divergence between $\mu_\star$ and μ_θ when $n \to \infty$ (see e.g. Müller, 2013), and it might be that the WABC posterior inherits similar properties for faster decaying thresholds.

In high dimensions, the rate of convergence of the Wasserstein distance between empirical measures is known to be slow (Talagrand, 1994). On the other hand, recent results establish that it concentrates quickly around its expectation: For instance, del Barrio and Loubes (2017) show that regardless of dimension, $\mathfrak{W}_2^2(\hat{\mu}_n, \hat{\mu}_{\theta,n}) - \mathbb{E}\mathfrak{W}_2^2(\hat{\mu}_n, \hat{\mu}_{\theta,n})$ converges weakly at the $\sqrt{n}$ rate to a centered Gaussian random variable with known (finite) variance $\sigma^2(\mu_\star, \mu_\theta)$. If the map $\theta \mapsto \mathbb{E}\mathfrak{W}_2^2(\hat{\mu}_n, \hat{\mu}_{\theta,n})$ offers discrimination between the parameters that is similar to $\theta \mapsto \mathfrak{W}_2^2(\mu_\star, \mu_\theta)$, it is not clear how the Wasserstein distance's convergence rate would impact the WABC posterior. Detailed analysis of WABC's dependence on dimension is an interesting avenue of future research.

4. Time series

Viewing data sets as empirical distributions requires some additional care in the case of dependent data, which are common in settings where ABC methods are applied. A naïve approach consists in ignoring dependencies, which might be enough to estimate all parameters in some cases, as illustrated in Section 5.3. However, in general, ignoring dependencies might prevent some parameters from being identifiable, as illustrated in the examples of this section. We propose two main approaches to extend the WABC methodology to time series.

4.1. Curve matching

Visually, we might consider two time series to be similar if their curves are similar, in a trace plot of the series in the vertical axis against the time indices on the horizontal axis. The Euclidean vector distance between curves sums the vertical differences between pairs of points with identical time indices. We can instead introduce the points $\tilde{y}_t = (t, y_t)$ and $\tilde{z}_t = (t, z_t)$ for all $t \in 1:n$, viewing the trace plot as a scatter plot. The distance between two points, (t, y_t) and (s, z_s), can be measured by a weighted distance $\rho_\lambda\left((t, y_t), (s, z_s)\right) = \|y_t - z_s\| + \lambda|t - s|$, where λ is a non-negative weight, and $\|y - z\|$ refers to the Euclidean distance between y and z. Intuitively, the distance ρ_λ takes into account both vertical and horizontal differences between points of the curves, λ tuning the importance of horizontal differences relative to vertical differences. We can then define the Wasserstein distance between two empirical measures supported by $\tilde{y}_{1:n}$ and $\tilde{z}_{1:n}$, with ρ_λ as a ground distance on the observation space $\{1, \ldots, n\} \times \mathcal{Y}$. Since computing the Wasserstein distance can be thought of as solving an assignment problem, a large value of λ implies that y_t will be assigned to z_t, for all t. The transport cost will then be $n^{-1}\sum_{t=1}^n \|y_t - z_t\|$, corresponding to the Euclidean distance (up to a scaling factor). If λ is smaller, (t, y_t) is assigned to some (s, z_s), for some s possibly different than t. If λ goes to zero, the distance coincides with the Wasserstein distance between the marginal empirical distributions of $y_{1:n}$ and $z_{1:n}$, where the time element is entirely ignored. Thus curve matching provides a compromise between the Euclidean distance between the series seen as vectors, and the Wasserstein distance between marginal empirical distributions.

For any $\lambda > 0$, the curve matching distance satisfies condition (b) of Proposition 3.1, implying that the resulting WABC posterior converges to the standard posterior distribution as $\varepsilon \to 0$. To estimate the WABC posterior, we can utilize any of the methods

for computing and approximating the Wasserstein distance discussed in Section 2.3 in combination with the SMC algorithm of Section 2.1. In Example 4.1, we use the exact Wasserstein curve matching distance to infer parameters in a cosine model. The choice of λ is open, but a simple heuristic for univariate time series goes as follows. Consider the aspect ratio of the trace plot of the time series (y_t), with horizontal axis spanning from 1 to n, and vertical axis from $\min_{t\in 1:n} y_t$ to $\max_{t\in 1:n} y_t$. For an aspect ratio of $H : V$, one can choose λ as $((\max_{t\in 1:n} y_t - \min_{t\in 1:n} y_t)/V) \times (H/n)$. For this choice ρ_λ corresponds to the Euclidean distance in a rectangular plot with the given aspect ratio.

Generalizations of the curve matching distance have been proposed independently by Thorpe et al. (2017) under the name "transportation L_p distances". In that paper, the properties of the curve matching distance are studied in detail, and compared to and combined with the related notion of dynamic time warping (Berndt and Clifford, 1994). Other related distances between time series include the Skorokhod distance between curves (Majumdar and Prabhu, 2015) and the Fréchet distance between polygons (Buchin et al., 2008), in which y_t would be compared to $z_{r(t)}$, where r is a retiming function to be optimized.

EXAMPLE 4.1. *Consider a cosine model where* $y_t = A\cos(2\pi\omega t + \phi) + \sigma w_t$, *where* $w_t \sim \mathcal{N}(0,1)$, *for all* $t \geq 1$, *are independent. Information about* ω *and* ϕ *is mostly lost when considering the marginal empirical distribution of* $y_{1:n}$. *In Figure 2, we compare the ABC posteriors obtained either with the Euclidean distance between the series, or with curve matching, with an aspect ratio of one; in both cases the algorithm is run for* 10^6 *model simulations. The figure also shows an approximation of the exact posterior distribution, obtained via Metropolis–Hastings. The prior distributions are uniform on* $[0, 1/10]$ *and* $[0, 2\pi]$ *for* ω *and* ϕ *respectively, and standard Normal on* $\log(\sigma)$ *and* $\log(A)$. *The data are generated using* $\omega = 1/80$, $\phi = \pi/4$, $\log(\sigma) = 0$ *and* $\log(A) = \log(2)$, *with* $n = 100$. *We see that curve matching yields a more satisfactory estimation of* σ *in Figure 2c, and a similar approximation for the other parameters. By contrast, an ABC approach based on the marginal distribution of* $y_{1:n}$ *would fail to identify* ϕ.

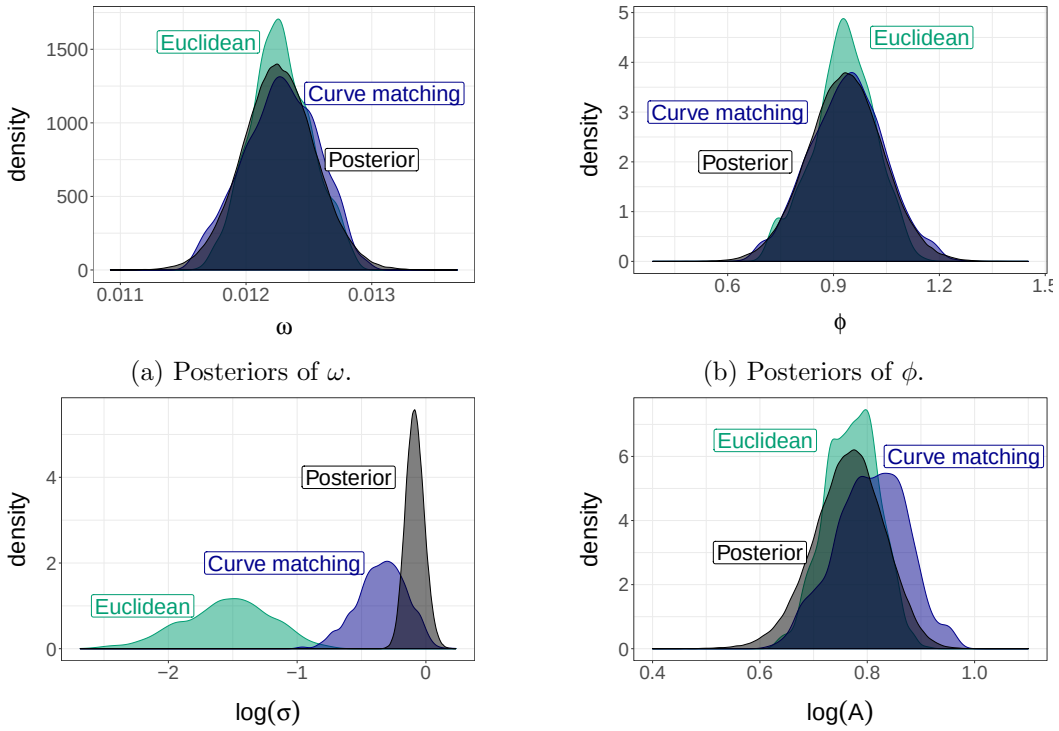
4.2. Reconstructions

Our second approach consists in transforming the time series to define an empirical distribution $\tilde{\mu}_n$ from which parameters can be estimated.

4.2.1. Delay reconstruction

In time series analysis, the lag-plot is a scatter plot of the pairs $(y_t, y_{t-k})_{t=k+1}^n$, for some lag $k \in \mathbb{N}$, from which one can inspect the dependencies between lagged values of the series. In ABC applied to time series models, lag-k autocovariances, defined as the sample covariance of $(y_t, y_{t-k})_{t=k+1}^n$, are commonly used statistics to summarize these dependencies (Marin et al., 2012; Mengersen et al., 2013; Li and Fearnhead, 2018). Here, we also propose to use the joint samples $(y_t, y_{t-k})_{t=k+1}^n$, but bypass their summarization into sample covariances. In particular, we define delay reconstructions as $\tilde{y}_t = (y_t, y_{t-\tau_1}, \ldots, y_{t-\tau_k})$ for some integers $\tau_1, \ldots, \tau_k$. The sequence, denoted $\tilde{y}_{1:n}$ after relabelling and redefining n, inherits many properties from the original series, such as stationarity. Therefore, the empirical distribution of $\tilde{y}_{1:n}$, denoted by $\tilde{\mu}_n$, might converge to a limit $\tilde{\mu}_\star$. In turn, $\tilde{\mu}_\star$ is likely to capture more features of the dependency structure than $\mu_\star$, and the resulting procedure might provide more accurate inference on the model parameters than if we were to compare the lag-k autocovariances alone.

Delay reconstructions (or embeddings) play a central role in dynamical systems (Kantz and Schreiber, 2004), for instance in Takens' theorem and variants thereof (Stark et al.,



(a) Posteriors of ω.

(b) Posteriors of ϕ.

(c) Posteriors of $\log(\sigma)$.

(d) Posteriors of $\log(A)$.

Fig. 2: ABC posterior samples in the cosine model of Example 4.1, using either the Euclidean distance or curve matching with the exact Wasserstein distance and $\lambda = 1$, after 10^6 model simulations. We compare to the posterior distribution, obtained using the 50,000 last samples in a Metropolis–Hastings chain of length 100,000. The standard deviation of the noise σ is better estimated with curve matching than with the Euclidean distance between time series.

2003). The Wasserstein distance between the empirical distributions of delay reconstructions has previously been proposed as a way of measuring distance between time series (Moeckel and Murray, 1997; Muskulus and Verduyn-Lunel, 2011), but not as a device for parameter inference. In the ABC setting, we propose to construct the delay reconstructions of each synthetic time series, and to compute the Wasserstein distance between their empirical distribution and the empirical distribution of $\tilde{y}_{1:n}$. We refer to this approach as WABC with delay reconstruction.

Denote by $\tilde{\mu}_{\theta,n}$ the empirical distribution of the delay reconstructed series $\tilde{z}_{1:n}$, formed from $z_{1:n} \sim \mu_\theta^{(n)}$. Then, provided that $\tilde{\mu}_{\theta,n}$ converges to an identifiable distribution $\tilde{\mu}_\theta$ as $n \to \infty$, we are back in a setting where we can study the concentration behavior of the WABC posterior around $\tilde{\theta}_\star = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\tilde{\mu}_\star, \tilde{\mu}_\theta)$, assuming existence and uniqueness (see Section 3.3). In well-specified settings, $\tilde{\theta}_\star$ must correspond to the data-generating parameters.

When the entries of the vectors $y_{1:n}$ and $z_{1:n}$ are all unique, which happens with probability one when $\mu_\star^{(n)}$ and $\mu_\theta^{(n)}$ are continuous distributions, then $\mathfrak{W}_p(\tilde{y}_{1:n}, \tilde{z}_{1:n}) = 0$ if and only if $y_{1:n} = z_{1:n}$. To see this, consider the setting where $\tilde{y}_t = (y_t, y_{t-1})$, and $\tilde{z}_t = (z_t, z_{t-1})$. For the empirical distributions of $\tilde{y}_{1:n}$ and $\tilde{z}_{1:n}$ to be equal, we require that for every t there exists a unique s such that $\tilde{y}_t = \tilde{z}_s$. However, since the values in $y_{1:n}$ and $z_{1:n}$ are unique, the values y_1 and z_1 appear only as the second coordinates of $\tilde{y}_2$ and $\tilde{z}_2$ respectively. It therefore has to be that $y_1 = z_1$ and $\tilde{y}_2 = \tilde{z}_2$. In turn, this implies that $y_2 = z_2$, and inductively, $y_t = z_t$ for all $t \in 1 : n$. A similar reasoning can be done for

any $k \geq 2$ and $1 \leq \tau_1 < \ldots < \tau_k$. This property can be used to establish the convergence of the WABC posterior based on delay reconstruction to the posterior, as $\varepsilon \to 0$, which can be deduced using condition (b) of Proposition 3.1.

In practice, for a non-zero value of ε, the obtained ABC posteriors might be different from the posterior, but still identify the parameters with a reasonable accuracy, as illustrated in Example 4.2. The quality of the approximation will depend on the choice of lags $\tau_1, \ldots, \tau_k$. Data-driven ways of making such choices are discussed by Kantz and Schreiber (2004). Still, since the order of the original data is only partly reflected in delay reconstructions, some model parameters might be difficult to estimate with delay reconstruction, such as the phase shift ϕ in Example 4.1.

EXAMPLE 4.2. *Consider an autoregressive process of order 1, written AR(1), where* $y_1 \sim \mathcal{N}(0, \sigma^2/(1-\phi^2))$, *for some* $\sigma > 0$ *and* $\phi \in (-1, 1)$. *For each* $t \geq 2$, *let* $y_t = \phi y_{t-1} + \sigma w_t$, *where* $w_t \sim \mathcal{N}(0,1)$ *are independent. The marginal distribution of each* y_t *is* $\mathcal{N}(0, \sigma^2/(1-\phi^2))$. *Furthermore, by an ergodic theorem, the empirical distribution* $\hat{\mu}_n$ *of the time series converges to this marginal distribution. The two parameters* (ϕ, σ^2) *are not identifiable from the limit* $\mathcal{N}(0, \sigma^2/(1-\phi^2))$. *Figure 3a shows WABC posterior samples derived while ignoring time dependence, obtained for decreasing values of* ε, *using a budget of* 10^5 *model simulations. The prior is uniform on* $[-1, 1]$ *for* ϕ, *and standard Normal on* $\log(\sigma)$. *The data are generated using* $\phi = 0.7, \log(\sigma) = 0.9$ *and* $n = 1,000$. *The WABC posteriors concentrate on a ridge of values with constant* $\sigma^2/(1-\phi^2)$.

Using $k = 1$, *we consider* $\tilde{y}_t = (y_t, y_{t-1})$ *for* $t \geq 2$. *The reconstructions are then sub-sampled to* 500 *values,* $\tilde{y}_2 = (y_2, y_1), \tilde{y}_4 = (y_4, y_3), \ldots, \tilde{y}_{1000} = (y_{1000}, y_{999})$; *similar results were obtained with the* 999 *reconstructed values, but sub-sampling leads to computational gains in the exact Wasserstein distance calculations; see Section 2.3.4. The stationary distribution of* $\tilde{y}_t$ *is given by*

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \frac{\sigma^2}{1-\phi^2}\begin{pmatrix}1 & \phi\\ \phi & 1\end{pmatrix}\right). \tag{8}$$

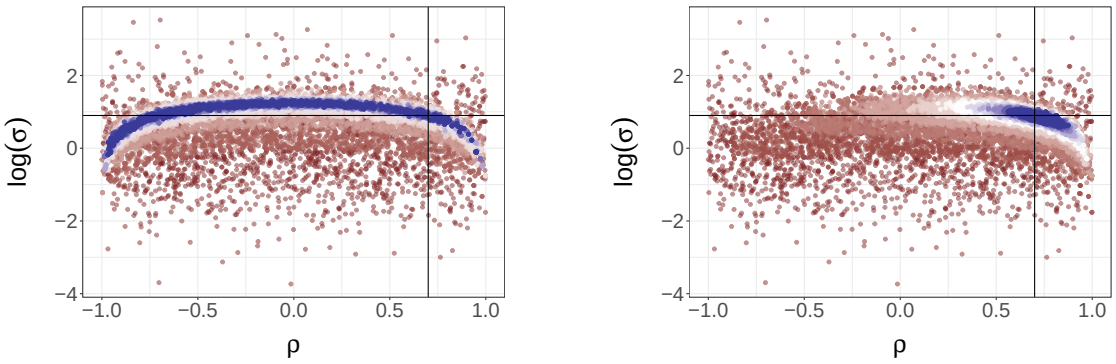
Both parameters σ^2 *and* ϕ *can be identified from a sample approximating the above distribution. Figure 3b shows the WABC posteriors obtained with delay reconstruction and a budged of* 10^5 *model simulations concentrating around the data-generating values as* ε *decreases.*

4.2.2. Residual reconstruction

Another approach to handle dependent data is advocated in Mengersen et al. (2013), in the context of ABC via empirical likelihood. In various time series models, the observations are modeled as transformations of some parameter θ and residual variables $w_1, \ldots, w_n$. Then, given a parameter θ, one might be able to reconstruct the residuals corresponding to the observations. In Example 4.1, one can define $w_t = (y_t - A\cos(2\pi\omega t + \phi))/\sigma$. In Example 4.2, one can define $w_t = (y_t - \phi y_{t-1})/\sigma$; other examples are given in Mengersen et al. (2013). Once the residuals have been reconstructed, their empirical distribution can be compared to the distribution that they would follow under the model, e.g. a standard Normal in Examples 4.2 and 4.1.

5. Numerical experiments

We illustrate the proposed approach and make comparisons to existing methods in various models taken from the literature. In each example, we approximate the WABC posterior using the SMC algorithm and default parameters outlined in Section 2.1.



(a) WABC using the Wasserstein distance between marginal distributions.

(b) WABC using the Wasserstein distance between empirical distributions of delay reconstructions.

Fig. 3: Samples from the WABC posteriors of $(\phi, \log(\sigma))$ in the AR(1) model of Example 4.2, as ε decreases over the steps of the SMC sampler (colors from red to white to blue). On the left, using the marginal empirical distribution of the series, the WABC posteriors concentrate around a ridge of values such that $\sigma^2/(1-\phi^2)$ is constant. On the right, using delay reconstruction with lag $k = 1$, the WABC posteriors concentrate around the data-generating parameters, $\phi = 0.7, \log(\sigma) = 0.9$, indicated by full lines. Both methods had a total budget of 10^5 model simulations.

5.1. Quantile "g-and-k" distribution

We first consider an example where the likelihood can be approximated to high precision, which allows comparisons between the standard posterior and WABC approximations. We observe that the WABC posterior converges to the true posterior in the univariate g-and-k example, as suggested by Proposition 3.1. We also compare WABC to a method developed by Fearnhead and Prangle (2012) that uses a semi-automatic construction of summary statistics. Lastly, we compare the use of the Wasserstein distance with other distances described in Section 2.3 on a bivariate version of the g-and-k distribution.

5.1.1. Univariate "g-and-k"

A classical example in the ABC literature (see e.g. Fearnhead and Prangle, 2012; Mengersen et al., 2013), the univariate g-and-k distribution is defined in terms of its quantile function:

$$r \in (0,1) \mapsto a + b\left(1 + 0.8\frac{1-\exp(-gz(r))}{1+\exp(-gz(r))}\right)\left(1+z(r)^2\right)^k z(r), \tag{9}$$

where $z(r)$ refers to the r-th quantile of the standard Normal distribution.

Sampling from the g-and-k distribution can be done by plugging standard Normal variables into (9) in place of $z(r)$. The probability density function is intractable, but can be numerically calculated with high precision since it only involves one-dimensional inversions and differentiations of the quantile function in (9), as described in Rayner and MacGillivray (2002). Therefore, Bayesian inference can be carried out with e.g. Markov chain Monte Carlo.

We generate $n = 250$ observations from the model using $a = 3, b = 1, g = 2, k = 0.5$, and the parameters are assigned a uniform prior on $[0, 10]^4$. We estimate the posterior distribution by running 5 Metropolis–Hastings chains for $75,000$ iterations, and discard the first $50,000$ as burn-in. For the WABC approximation, we use the SMC sampler outlined in Section 2.1 with $N = 2,048$ particles, for a total of 2.4×10^6 simulations from the model. The resulting marginal WABC posteriors are also compared to the marginal

posteriors obtained with the semi-automatic ABC approach of Fearnhead and Prangle (2012). We used the rejection sampler in the `abctools` package (Nunes and Prangle, 2015), also for a total of 2.4×10^6 model simulations, of which $N = 2,048$ draws from the prior are accepted. We observed no benefit to accepting fewer draws. The semi-automatic approach requires the user to specify a set of initial summary statistics, for which we used every 25th order statistic as well as the minimum (that is, $y_{(1)}, y_{(25)}, y_{(50)}, \ldots, y_{(250)}$) and their powers up to fourth order, following guidance in Fearnhead and Prangle (2012).

Figure 4 shows the marginal posterior distributions and their approximations obtained with WABC and semi-automatic ABC. The plots show that the WABC posteriors appear to be closer to the target distributions, especially on the a, b and k parameters. Neither method captures the marginal posterior of g well, though the WABC posterior appears more concentrated in the region of interest on that parameter as well.

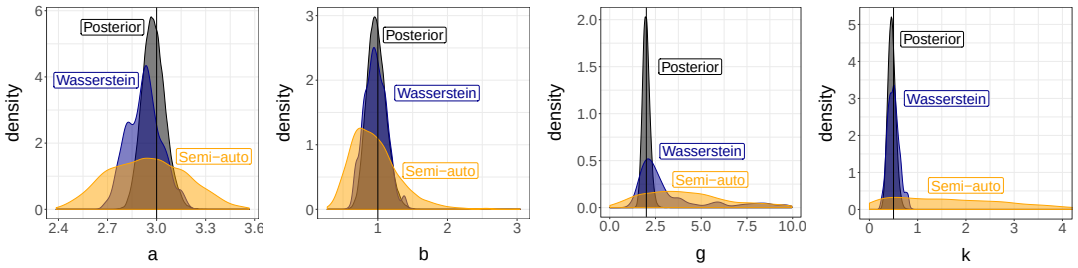
In both of the ABC approaches, the main computational costs stem from simulating from the model and sorting the resulting data sets. Over 1,000 repetitions, the average wall-clock time to simulate a data set was $7.7 \times 10^{-5}s$ on an Intel Core i5 (2.5GHz). The average time to sort the resulting data sets was $8.9 \times 10^{-5}s$, and computing the Wasserstein distance to the observed data set was negligibly different from this. In semi-automatic ABC, one additionally has to perform a regression step. The rejection sampler in semi-automatic ABC is easier to parallelize than our SMC approach, but on the other hand requires more memory due to the regression used in constructing the summary statistic. This makes the method hard to scale up beyond the number of model simulations considered here, without using specialized tools for large scale regression.

This problem does not arise in the WABC sequential Monte Carlo sampler, and Figure 5 illustrates the behavior of the marginal WABC posteriors as more steps of the SMC sampler are performed. In particular, we can see that the WABC approximations for a, b and k converge to the corresponding posteriors (up to some noise). The approximations for g also shows convergence towards the posterior, but have not yet reached the target distribution at the stage when the sampler was terminated. The convergence is further illustrated in Figure 5g, where the $\mathfrak{W}_1$- distance between the joint WABC posterior and joint posterior is plotted as a function of the number of simulations from the model. The plot shows that the Wasserstein distance between the distributions decreases from around 10 to around 0.06 over the course of the SMC run. The distances are approximated by thinning the MCMC samples left after burn-in down to 2,048 samples, and computing the Wasserstein between the corresponding empirical distribution and the empirical distributions supported at the $N = 2,048$ SMC particles at each step.

Figure 5f shows the development of the threshold as a function of the number of model simulations, showing that ε decreases to 0.07 over the course of the SMC. The threshold decreases at a slower rate as it approaches zero, suggesting that the underlying sampling problem becomes harder as ε becomes smaller. This is also illustrated by Figure 5e, which shows the number of model simulations performed at each step of the SMC algorithm. This number is increasing throughout the run of the algorithm, as the r-hit kernel requires more and more attempts before it reaches the desired number of hits.

5.1.2. Bivariate "g-and-k"

We also consider the bivariate extension of the g-and-k distribution (Drovandi and Pettitt, 2011), where one generates bivariate Normals with mean zero, variance one, and correlation ρ, and substitutes $z(r)$ with them in (9), with parameters (a_i, b_i, g_i, k_i) for each component $i \in \{1, 2\}$. Since the model generates bivariate data, we can no longer rely on simple sorting to calculate the Wasserstein distance. We compare the exact Wasserstein distance to the approximations discussed in Section 2.3, as well as the maximum mean discrepancy, whose use within ABC was proposed by Park et al. (2016) (see Section 1.4).



(a) Posteriors of a. (b) Posteriors of b. (c) Posteriors of g. (d) Posteriors of k.

Fig. 4: Posterior marginals in the univariate g-and-k example of Section 5.1.1 (obtained via MCMC), approximations by Wasserstein ABC and semi-automatic ABC, each with a budget of 2.4×10^6 model simulations. Data-generating values are indicated by vertical lines.

We generate $n = 500$ observations from the model using $a_1 = 3, b_1 = 1, g_1 = 1, k_1 = 0.5, a_2 = 4, b_2 = 0.5, g_2 = 2, k_2 = 0.4, \rho = 0.6$, as in Section 5.2 of Drovandi and Pettitt (2011). The parameters (a_i, b_i, g_i, k_i) are assigned a uniform prior on $[0, 10]^4$, independently for $i \in \{1, 2\}$, and ρ a uniform prior on $[-1, 1]$. We estimate the posterior distribution by running 8 Metropolis–Hastings chains for $150,000$ iterations, and discard the first $50,000$ as burn-in. For each of the ABC approximations, we run the SMC sampler outlined in Section 2.1 for a total of 2×10^6 simulations from the model. For the MMD, we use the estimator

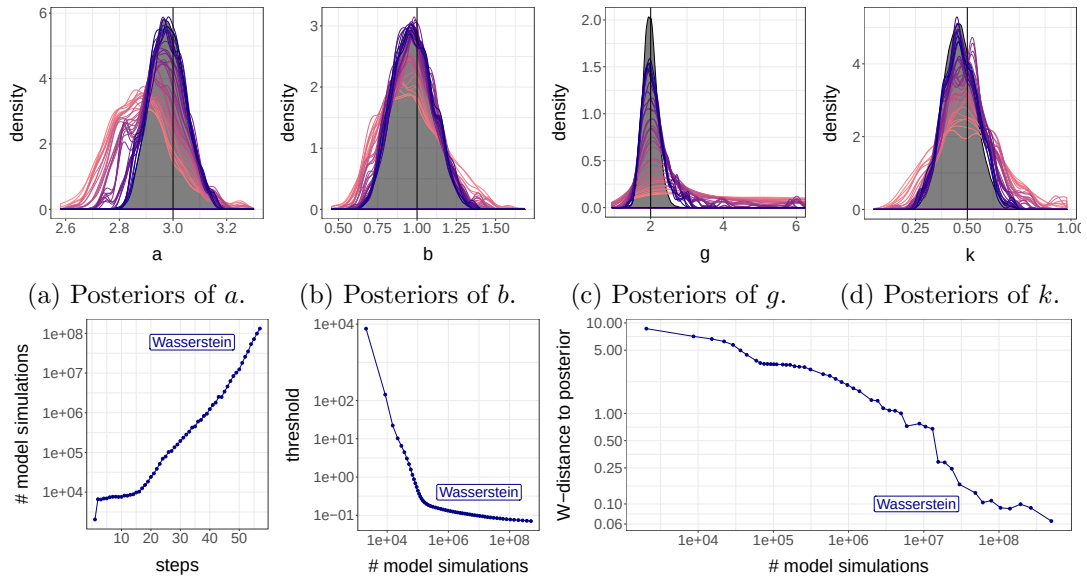
$$\mathrm{MMD}^2(y_{1:n}, z_{1:n}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(z_i, z_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(y_i, z_j), \tag{10}$$

with the kernel $k(x, x') = \exp\left(-\|x - x'\|^2/2h^2\right)$. The bandwidth h was fixed to be the median of the set $\{\|y_i - y_j\|_1 : i, j = 1, \ldots, n\}$, following guidance in Park et al. (2016).

Figure 6c shows the $\mathfrak{W}_1$-distance between the $N = 2,048$ ABC posterior samples, obtained with the different distances, and a sample of $2,048$ points thinned out from the Markov chains targeting the posterior. This distance is plotted against the number of model simulations. It shows that all distances yield ABC posteriors that get closer to the actual posterior. On the other hand, for this number of model simulations, all of the ABC posteriors are significantly different from the actual posterior. For comparison, the $\mathfrak{W}_1$-distance between two samples of size $2,048$ thinned out from the Markov chains is on average about 0.07. In this example, it appears that the MMD leads to ABC posteriors that are not as close to the posterior as the other distances given the same budget of model simulations. The Hilbert distance provides a particularly cheap and efficient alternative to the Wasserstein distance in this bivariate case, providing a very similar approximation to the posterior as both the exact Wasserstein and swapping distances.

Figure 6b shows the development of the threshold as a function of the number of model simulations for the different distances. Note that the MMD is not on the same scale as the Wasserstein distance and its approximations, and therefore the MMD thresholds are not directly comparable to the those of the other distances. As for the univariate g-and-k distribution, the thresholds decrease at a slower rate as they become smaller for each of the distances, suggesting that the underlying sampling problem becomes harder as ε becomes smaller. The behaviors of the thresholds based on the exact Wasserstein, swapping, and Hilbert distances appear negligibly different. Figure 6a shows the number of model simulations performed at each step of the SMC algorithm for each of the distances. As before, these numbers are increasing throughout the run of the algorithm, as the r-hit kernel requires more and more attempts before it reaches the desired number of hits.

An important distinction between the distances is the time they take to compute. For



(a) Posteriors of a. (b) Posteriors of b. (c) Posteriors of g. (d) Posteriors of k.

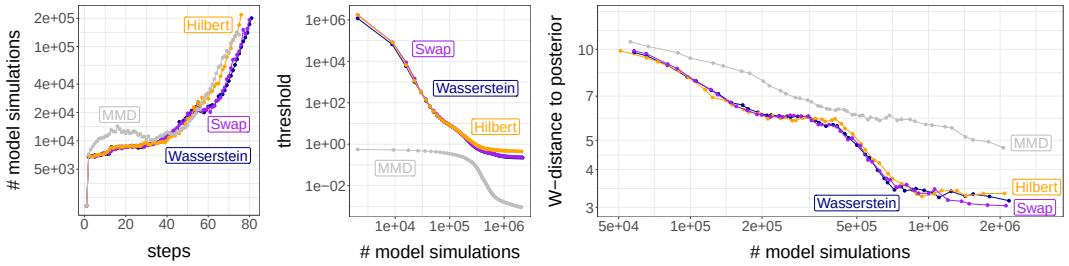
(e) Model simulations vs. SMC steps. (f) Threshold ε vs. model simulations (g) $\mathfrak{W}_1$-distance to posterior vs. number of model simulations.

Fig. 5: 5a-5d: Posterior marginals in the univariate g-and-k example of Section 5.1.1 (grey, obtained via MCMC) and approximations by Wasserstein ABC from step 20 to step 57 of the SMC algorithm. The color of the WABC approximation changes from red to blue as more steps of the SMC sampler are performed, decreasing the threshold ε. For the densities plotted here, the threshold reduces from $\varepsilon = 0.20$ to $\varepsilon = 0.07$. The range of the plot for g has been truncated to $(1, 6)$ to highlight the region of interest, despite the densities from the earlier steps of the SMC having support outside this region. Data-generating values are indicated by vertical lines. Figure 5e shows the number of simulations from the model used in each step of the SMC algorithm (y-axis in log scale). This number is increasing due to use of the r-hit kernel within the SMC. Figures 5f and 5g show the threshold ε and the $\mathfrak{W}_1$-distance to the posterior respectively, against the number of model simulations (both plots in log-log scale).

data sets of size $n = 500$ simulated using the data-generating parameter, the average wall-clock times to compute distances between simulated and observed data, on an Intel Core i7-5820K (3.30GHz), are as follows: $0.002s$ for the Hilbert distance, $0.01s$ for the MMD, $0.03s$ for the swapping distance, and $0.22s$ for the exact Wasserstein distance; these average times were computed on $1,000$ independent data sets. In this example, simulating from the model takes a negligible amount of time, even compared to the Hilbert distance. Calculating the likelihood over $1,000$ parameters drawn from the prior, we find an average computing time of $0.05s$. In combination with the information conveyed by Figure 6, the Hilbert and swapping-based ABC posteriors provide good approximations of the exact Wasserstein-based ABC posteriors in only a fraction of the time the latter takes to compute.

5.2. Toggle switch model

We borrow the system biology "toggle switch" model used in Bonassi et al. (2011); Bonassi and West (2015), inspired by studies of dynamic cellular networks. This provides an example where the design of specialized summaries can be replaced by the Wasserstein distance between empirical distributions. For $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, let $(u_{i,t}, v_{i,t})$ denote the expression levels of two genes in cell i at time t. Starting from



(a) Model simulations vs. SMC steps. (b) Threshold ε vs. model simulations (c) $\mathfrak{W}_1$-distance to posterior vs. number of model simulations.

Fig. 6: 6a shows the number of simulations from the model used in each step of the SMC algorithm for the four distances applied to the bivariate g-and-k model of Section 5.1.2 (y-axis in log scale). This number is increasing due to use of the r-hit kernel within the SMC. Figure 6b shows the thresholds ε against the number of model simulations (in log-log scale). Note that the MMD is not on the same scale as the Wasserstein distance and its approximations, and therefore the MMD thresholds are not directly comparable to the those of the other distances. Figure 6c shows the $\mathfrak{W}_1$-distance between the joint ABC posteriors based on the different distances to the joint true posterior, against the number of model simulations (in log-log scale).

$(u_{i,0}, v_{i,0}) = (10, 10)$, the evolution of $(u_{i,t}, v_{i,t})$ is given by

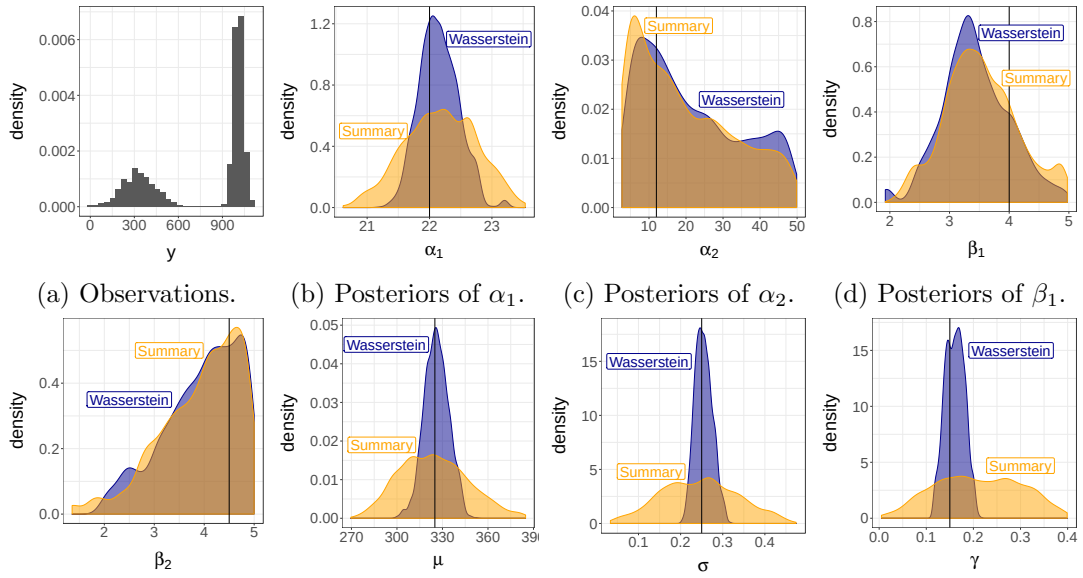
$$u_{i,t+1} = u_{i,t} + \alpha_1/(1 + v_{i,t}^{\beta_1}) - (1 + 0.03u_{i,t}) + 0.5\xi_{i,1,t},$$
$$v_{i,t+1} = v_{i,t} + \alpha_2/(1 + u_{i,t}^{\beta_2}) - (1 + 0.03v_{i,t}) + 0.5\xi_{i,2,t},$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are parameters, and ξ's are standard Normal variables, truncated so that $(u_{i,t}, v_{i,t})$ only takes non-negative values. For each cell i, we only observe a noisy measurement of the terminal expression level $u_{i,T}$. Specifically, the observations y_i are assumed to be independently distributed as $\mathcal{N}(\mu + u_{i,T}, \mu^2\sigma^2/u_{i,T}^{2\gamma})$ random variables truncated to be non-negative, where μ, σ, γ are parameters. We generate $n = 2,000$ observations using $\alpha_1 = 22$, $\alpha_2 = 12$, $\beta_1 = 4$, $\beta_2 = 4.5$, $\mu = 325$, $\sigma = 0.25$, $\gamma = 0.15$. A histogram of the data is shown in Figure 7a.

We consider the task of estimating the data-generating values, using uniform prior distributions on $[0, 50]$ for α_1, α_2, on $[0, 5]$ for β_1, β_2, on $[250, 450]$ for μ, $[0, 0.5]$ for σ and on $[0, 0.4]$ for γ. These ranges are derived from Figure 5 in Bonassi and West (2015). We compare our method using $p = 1$ with a summary-based approach using the 11-dimensional tailor-made summary statistic from Bonassi et al. (2011); Bonassi and West (2015). Since the data are one-dimensional, the Wasserstein distance between data sets can be computed quickly via sorting. For both methods, we use the SMC sampler outlined in Section 2.1, for a total number of 10^6 model simulations.

The seven marginal ABC posterior distributions obtained in the final step of the SMC sampler are shown in Figure 7. We find that the marginal WABC and summary-based posteriors concentrate to the same distributions for the α_2, β_1 and β_2 parameters. On the remaining parameters, the marginal WABC posteriors show stronger concentration around the data-generating parameters than the summary-based approach. Comparing the results, we see that the design of a custom summary can be bypassed using the Wasserstein distance between empirical distributions: the resulting posterior approximations appear to be more concentrated around the data-generating parameters, and our proposed approach is fully black-box. The time to simulate data from the model does not seem to depend noticeably on the parameter, and the average wall-clock time to simulate a data set over 1,000 repetitions was $0.523s$ on an Intel Core i5 (2.5GHz). The average time compute the Wasserstein distance to the observed data set was $0.0002s$, whereas the

average time to compute the summary statistic was $0.176s$.



(a) Observations. (b) Posteriors of α_1. (c) Posteriors of α_2. (d) Posteriors of β_1.

(e) Posteriors of β_2. (f) Posteriors of μ. (g) Posteriors of σ. (h) Posteriors of γ.

Fig. 7: Histogram of observations (7a), and marginal posteriors based on WABC and the summary statistic from Bonassi et al. (2011); Bonassi and West (2015) in the toggle switch model. The ABC posteriors are computed using the SMC sampler from Section 2.1, for a total number of 10^6 model simulations. Data-generating values are indicated by vertical lines.

5.3. Queueing model

We turn to the M/G/1 queueing model, which has appeared frequently as a test case in the ABC literature, see e.g. Fearnhead and Prangle (2012). It provides an example where the observations are dependent, but where the parameters can be identified from the marginal distribution of the data. In the model, customers arrive at a server with independent interarrival times w_i, exponentially distributed with rate θ_3. Each customer is served with independent service times u_i, taken to be uniformly distributed on $[\theta_1, \theta_2]$. We observe only the interdeparture times y_i, given by the process $y_i = u_i + \max\{0, \sum_{j=1}^{i} w_j - \sum_{j=1}^{i-1} y_j\}$. The prior on $(\theta_1, \theta_2 - \theta_1, \theta_3)$ is Uniform on $[0, 10]^2 \times [0, 1/3]$.

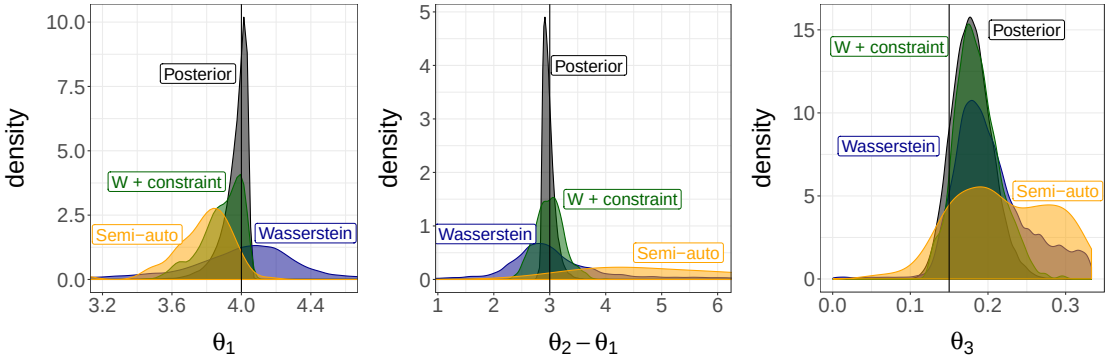
We use the data set given in Shestopaloff and Neal (2014), which was generated using the parameters $(\theta_1, \theta_2 - \theta_1, \theta_3) = (4, 3, 0.15)$ and $n = 50$. The WABC posterior based on the empirical distribution of $y_{1:n}$, ignoring dependencies, is approximated using the SMC algorithm of Section 2.1, with a budget of 10^7 model simulations. We compare with the semi-automatic ABC approach of Fearnhead and Prangle (2012) with the same budget of model simulations, using a subset of 20 evenly spaced order statistics as the initial summary statistics in that method. The semi-automatic ABC posteriors are computed using the rejection sampler in the `abctools` package (Nunes and Prangle, 2015), accepting the 100 best samples. The actual posterior distribution is approximated with a particle marginal Metropolis–Hastings (PMMH) run (Andrieu et al., 2010), using 4,096 particles and 10^5 iterations. The use of PMMH was suggested in Shestopaloff and Neal (2014) as an alternative to the model-specific Markov chain Monte Carlo algorithm they propose.

Upon observing $y_{1:n}$, θ_1 has to be less than $\min_{i \in 1:n} y_i$, which is implicitly encoded in the likelihood, but not in an ABC procedure. One can add this constraint explicitly, rejecting parameters that violate it, which is equivalent to redefining the prior on θ_1 to be

uniform on $[0, \min_{i\in 1:n} y_i]$. Figure 8 shows the marginal distributions of the parameters obtained with PMMH, semi-automatic ABC, and WABC, with or without the additional constraint.

Overall, the WABC approximations are close to the posterior, in comparison to the relatively vague prior distribution on $(\theta_1, \theta_2 - \theta_1)$. Furthermore, we see that incorporating the constraint leads to marginal WABC approximations that are closer to the marginal posteriors. Both variations of WABC appear to perform better than semi-automatic ABC, except on θ_1, where the semi-automatic ABC approximation is closer to the posterior than the unconstrained WABC approximation. We observed no significant difference in the semi-automatic ABC posterior when incorporating the constraint on θ_1, and hence only show the approximated posterior for the unconstrained approach.

As in the univariate g-and-k model of Section 5.1.1, the computation costs for the WABC and semi-automatic approaches are similar, as they both rely on simulating from the model and sorting the resulting data. Over 1,000 repetitions, the average wall-clock time to simulate a data set was $7.5 \times 10^{-5} s$ on an Intel Core i5 (2.5GHz). Sorting a data set took on average $7.7 \times 10^{-5} s$, and computing the Wasserstein distance was negligibly different from this. For the semi-automatic ABC approach, one additionally has to perform the regression step. The model simulations in semi-automatic ABC are easier to parallelize, but the method is hard to scale up without specialized tools for large-scale regression, due to memory requirements of the regression used to construct the summary statistics.



(a) Posteriors of θ_1. (b) Posteriors of $\theta_2 - \theta_1$. (c) Posteriors of θ_3.

Fig. 8: Posterior marginals in the M/G/1 queueing model of Section 5.3 (obtained via particle marginal Metropolis–Hastings), approximations by Wasserstein ABC and semi-automatic ABC, and Wasserstein ABC accounting for the constraint that θ_1 has to be less than $\min_{i\in 1:n} y_i$, each with a budget of 10^7 model simulations. Data-generating values are indicated by vertical lines.

5.4. Lévy-driven stochastic volatility model

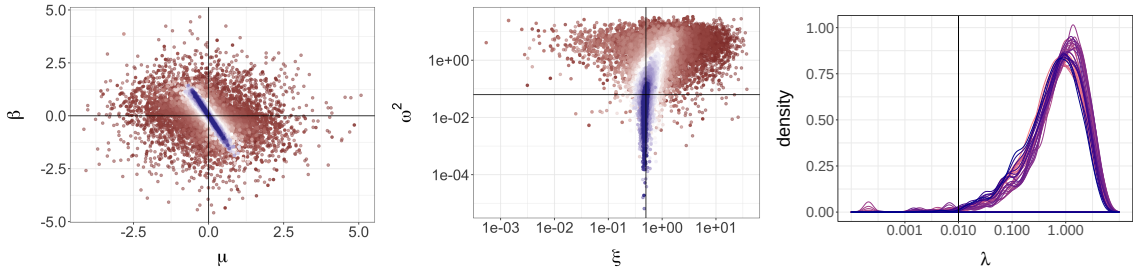
We consider a Lévy-driven stochastic volatility model (e.g. Barndorff-Nielsen and Shephard, 2002), used in Chopin et al. (2013) as a challenging example of parameter inference in state space models. We demonstrate how ABC with transport distances can identify some of the parameters in a black-box fashion, and can be combined with summaries to identify the remaining parameters. The observation y_t at time t is the log-return of a financial asset, assumed Normal with mean $\mu + \beta v_t$ and variance v_t, where v_t is the actual volatility. Together with the spot volatility z_t, the pair (v_t, z_t) constitutes a latent Markov chain, assumed to follow a Lévy process. Starting with $z_0 \sim \Gamma(\xi^2/\omega^2, \xi/\omega^2)$ (where the second parameter is the rate), and an arbitrary v_0, the evolution of the process goes as

follows:

$$k \sim \mathcal{P}\text{oisson}\left(\lambda\xi^2/\omega^2\right), \quad c_{1:k} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(t, t+1), \quad e_{1:k} \overset{\text{i.i.d.}}{\sim} \mathcal{E}\text{xp}\left(\xi/\omega^2\right),$$
$$z_{t+1} = e^{-\lambda} z_t + \sum_{j=1}^{k} e^{-\lambda(t+1-c_j)} e_j, \quad v_{t+1} = \frac{1}{\lambda}[z_t - z_{t+1} + \sum_{j=1}^{k} e_j]. \tag{11}$$

The random variables $(k, c_{1:k}, e_{1:k})$ are generated independently for each time period, and $1:k$ is the empty set when $k = 0$. The parameters are $(\mu, \beta, \xi, \omega^2, \lambda)$. We specify the prior as Normal with mean zero and variance 2 for μ and β, Exponential with rate 0.2 for ξ and ω^2, and Exponential with rate 1 for λ.

We generate synthetic data with $\mu = 0$, $\beta = 0$, $\xi = 0.5$, $\omega^2 = 0.0625$, $\lambda = 0.01$, which were used also in the simulation study of Barndorff-Nielsen and Shephard (2002); Chopin et al. (2013), of length $n = 10,000$. We use delay reconstruction with a lag $k = 1$, and the Hilbert distance $\mathfrak{H}_p$ of Section 2.3.2 with $p = 1$. Given the length of the time series, the cost of computing the Hilbert distance is much smaller than that of the other distances discussed in Section 2.3. We ran the SMC algorithm outlined in Section 2.1 until a total of 4.2×10^5 data sets had been simulated. Figure 9 shows the resulting quasi-posterior marginals for (μ, β), (ξ, ω^2), and λ. The parameters $(\mu, \beta, \xi, \omega^2)$ are accurately identified, from a vague prior to a region close to the data-generating values. On the other hand, the approximation of λ is barely different from the prior distribution. Indeed, the parameter λ represents a discount rate which impacts the long-range dependencies of the process, and is thus not captured by the bivariate marginal distribution of (y_t, y_{t-1}).



(a) Posteriors of (μ, β).

(b) Posteriors of (ξ, ω^2) (in log-log scale).

(c) Posteriors of λ (in log scale).

Fig. 9: ABC approximations in the Lévy-driven stochastic volatility model, using the Hilbert distance between delay reconstructions with lag $k = 1$. The plots show samples from the bivariate marginals of (μ, β) (left), (ξ, ω^2) (middle), and the marginal distributions of λ (right), as the threshold ε decreases during the steps of the SMC sampler (colors from red to blue). The total budget was 4.2×10^5 model simulations. Data-generating parameters are indicated by full lines.

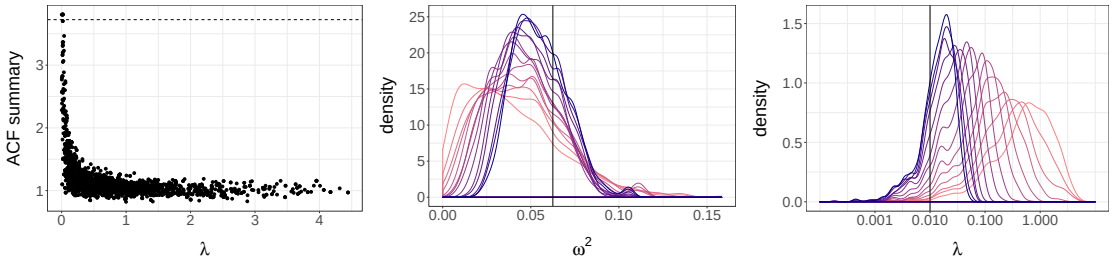
Hoping to capture long-range dependencies in the series, we define a summary $\eta(y_{1:n})$ as the sum of the first 50 sample autocorrelations among the squared observations. For each of the parameters obtained with the first run of WABC described above, we compute the summary of the associated synthetic data set. We plot the summaries against λ in Figure 10a. The dashed line indicates the value of the summary calculated on the observed data. The plot shows that the summaries closest to the observed summary are those obtained with the smallest values of λ. Therefore, we might be able to learn more about λ by combining the previous Hilbert distance with a distance between summaries.

Denote by $\mathfrak{H}_1(\tilde{y}_{1:n}, \tilde{z}_{1:n})$ the Hilbert distance between delay reconstructions, and by ε_h the threshold obtained after the first run of the algorithm. A new distance between data sets is defined as $|\eta(y_{1:n}) - \eta(z_{1:n})|$ if $\mathfrak{H}_1(\tilde{y}_{1:n}, \tilde{z}_{1:n}) < \varepsilon_h$, and $+\infty$ otherwise. We then run the SMC sampler of Section 2.1, initializing with the results of the first run, but using the

new distance. In this second run, a new threshold is introduced and adaptively decreased, keeping the first threshold ε_h fixed. One could also decrease both thresholds together or alternate between decreasing either. Note that the Hilbert distance and the summaries could have been combined in other ways, for instance in a weighted average.

We ran the algorithm with the new distance for an extra 6.6×10^5 model simulations. Figures 10b and 10c show the evolution of the WABC posterior distributions of ω^2 and λ during the second run. The WABC posteriors concentrate closer to the data-generating values, particularly for λ; for (μ, β, ξ), the effect is minimal and not shown. In terms of computing time, it took on average $1.3 \times 10^{-1} s$ to generate time series given the data-generating parameter, $2.4 \times 10^{-2} s$ to compute the Hilbert distance, and $1.5 \times 10^{-3} s$ to compute the summary statistic, on an Intel Core i5 (2.5GHz). Thus, most of the time consumed by the algorithm was spent generating data.

The WABC posterior could then be used to initialize a particle MCMC algorithm (Andrieu et al., 2010) targeting the posterior. The computational budget of roughly 1.1×10^6 model simulations, as performed in total by the WABC procedure in this section, would be equivalent to relatively few iterations of particle MCMC in terms of number of model transitions. Therefore, the cost of initializing a particle MCMC algorithm with the proposed ABC approach is likely to be negligible. The approach could be valuable in settings where it is difficult to initialize particle MCMC algorithms, for instance due to the large variance of the likelihood estimator for parameters located away from the posterior mode, as illustrated in Figure 2 (c) of Murray et al. (2013).



(a) Summary against λ. (b) Distributions of ω^2. (c) Distributions of λ (log-scale).

Fig. 10: Left: summary, defined as the sum of the first 50 sample autocorrelations of the squared series, against λ, computed for the output of the WABC algorithm using the Hilbert distance between delay reconstructions, applied to the Lévy-driven stochastic volatility model of Section 5.4. Middle and right: approximations of ω^2 and λ, from the second run of WABC using the Hilbert distance between delay reconstructions combined with the summary for another 6.6×10^5 model simulations. The colors change from red to blue as more steps of the SMC sampler are performed. The horizontal axis in the right plot is in log-scale, and illustrates the concentration of the ABC posterior towards the data-generating value of λ.

6. Discussion

Using the Wasserstein distance in approximate Bayesian computation leads to a new way of inferring parameters in generative models, bypassing the choice of summaries. The approach can also be readily used for deterministic models. We have demonstrated how the proposed approach can identify high posterior density regions, in settings of both i.i.d. (Section 5.1) and dependent data (Section 5.3). In some examples the proposed approximations appear to be at least as close to the posterior distribution as those produced by state-of-the-art summary-based ABC. For instance, in the toggle switch model of Section 5.2, our black-box method obtained posterior approximations that are more concentrated

on the data-generating parameters than those obtained with sophisticated, case-specific summaries, while being computationally cheaper. Furthermore, we have shown how summaries and transport distances can be fruitfully combined in Section 5.4. There are various ways of combining distances in the ABC approach, which could deserve more research.

We have proposed multiple ways of defining empirical distributions of time series data, in order to identify model parameters. The proposed approaches have tuning parameters, such as λ in the curve matching approach of Section 4.1 or the lags in delay reconstruction in Section 4.2. The choice of these parameters has been commented on by Thorpe et al. (2017) in the case of curve matching, and by Muskulus and Verduyn-Lunel (2011); Stark et al. (2003); Kantz and Schreiber (2004) in the case of delay reconstructions. Making efficient choices of these parameters might be easier than choosing summary statistics. Further research could leverage e.g. the literature on Skorokhod distances for λ (Majumdar and Prabhu, 2015). The investigation of similar methods for the setting of spatial data would also be interesting.

We have established some theoretical properties of the WABC distribution, adding to the existing literature on asymptotic properties of ABC posteriors (Frazier et al., 2018; Li and Fearnhead, 2018). In particular, we have considered settings where the threshold ε goes to zero for a fixed set of observations, and where the number of observations n goes to infinity with a slowly decreasing threshold sequence ε_n. In the first case, we establish conditions under which the WABC posterior converges to the posterior, as illustrated empirically in Section 5.1. In the second case, our results show that under certain conditions, the WABC posterior can concentrate in different regions of the parameter space compared to the posterior. We also derive upper bounds on the concentration rates, which highlight the potential impact of the order p of the Wasserstein distance, of the dimension of the observation space, and of model misspecification. The dependence on dimension of the observation space could be a particularly interesting avenue of future research

In comparison with the asymptotic regime, less is known about the properties of ABC posteriors for fixed ε. Viewing the WABC posterior as a coarsened posterior (Miller and Dunson, 2018), one can justify its use in terms of robustness to model misspecification. On the other hand, ABC posteriors in general do not yield conservative statements about the posterior for a fixed threshold ε and data set $y_{1:n}$. For instance, Figure 2c shows that ABC posteriors can have little overlap with the posterior, despite having shown signs of concentration away from the prior distribution.

As Wasserstein distance calculations scale super-quadratically with the number of observations n, we have introduced a new distance based on the Hilbert space-filling curve, computable in order $n \log n$, which can be used to initialize a swapping distance with a cost of order n^2. We have derived some posterior concentration results for the ABC posterior distributions using the Hilbert and swapping distances, similarly to Proposition 3.2 obtained for the Wasserstein distance. Many other distances related to optimal transport could be used; we mentioned Park et al. (2016) who used the maximum mean discrepancy, and recently Genevay et al. (2018) consider Sinkhorn divergences, and Jiang et al. (2018) consider the Kullback–Leibler divergence. A thorough comparison between these different distances, none of which involve summary statistics, could be of interest to ABC practitioners.

Acknowledgements We are grateful to Marco Cuturi, Jeremy Heng, Guillaume Pouliot, Neil Shephard, and the anonymous reviewers for helpful feedback. Pierre E. Jacob gratefully acknowledges support by the National Science Foundation through grant DMS-1712872.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo (with discussion). *Journal of the Royal Statistical Society: Series B*, **72**, 357–385. 22, 25

Barndorff-Nielsen, O. E. and Shephard, N. (2002) Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B*, **64**, 253–280. 23, 24

del Barrio, E. and Loubes, J.-M. (2017) Central limit theorems for empirical transportation cost in general dimension. *Preprint arXiv:1705.01299*, Universidad de Valladolid. 13

Bassetti, F., Bodini, A. and Regazzini, E. (2006) On minimum Kantorovich distance estimators. *Statistics & probability letters*, **76**, 1298–1302. 4

Basu, A., Shioya, H. and Park, C. (2011) *Statistical inference: the minimum distance approach*. CRC Press. 4

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate bayesian computation in population genetics. *Genetics*, **162**, 2025–2035. 1

Berndt, D. J. and Clifford, J. (1994) Using dynamic time warping to find patterns in time series. In *Proceedings of Workshop on Knowledge Discovery in Databases* (eds. U. M. Fayyad and R. Uthurusamy), 359–370. Palo Alto, CA: AAAI. 14

Bernton, E., Jacob, P. E., Gerber, M. and Robert, C. P. (2017) Inference in generative models using the Wasserstein distance. *Preprint arXiv:1701.05146*, Harvard University. 4

Bonassi, F. V. and West, M. (2015) Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis*, **10**, 171–187. 20, 21, 22

Bonassi, F. V., You, L. and West, M. (2011) Bayesian learning from marginal data in bionetwork models. *Statistical applications in genetics and molecular biology*, **10**. 20, 21, 22

Bonneel, N., Rabin, J., Peyré, G. and Pfister, H. (2015) Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, **51**, 22–45. 9

Buchin, K., Buchin, M. and Wenk, C. (2008) Computing the Fréchet distance between simple polygons. *Computational Geometry*, **41**, 2–20. 14

Burkard, R., Dell'Amico, M. and Martello, S. (2009) *Assignment Problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics. 7

Chopin, N., Jacob, P. and Papaspiliopoulos, O. (2013) SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B*, **75**, 397–426. 23, 24

Cuturi, M. (2013) Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26* (eds. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), 2292–2300. Red Hook, NY: Curran Associates, Inc. 2, 7

Del Moral, P., Doucet, A. and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, **22**, 1009–1020. 5, 6

Drovandi, C. C. and Pettitt, A. N. (2011) Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, **55**, 2541–2556. 18, 19

Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, **74**, 419–474. 3, 4, 5, 9, 17, 18, 22

Filippi, S., Barnes, C. P., Cornebise, J. and Stumpf, M. P. (2013) On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical applications in genetics and molecular biology*, **12**, 87–107. 6

Fournier, N. and Guillin, A. (2015) On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, **162**, 707–738. 11

Frazier, D. T., Martin, G. M., Robert, C. P. and Rousseau, J. (2018) Asymptotic properties of approximate Bayesian computation. *Biometrika*, **105**, 593–607. 10, 26

Genevay, A., Peyre, G. and Cuturi, M. (2018) Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (eds. A. Storkey and F. Perez-Cruz), vol. 84 of *Proceedings of Machine Learning Research*, 1608–1617. Lanzarote, Canary Islands: PMLR. 26

Gerber, M. and Chopin, N. (2015) Sequential quasi-Monte Carlo. *Journal of the Royal Statistical Society: Series B*, **77**, 509–579. 2, 8

Gerber, M., Chopin, N. and Whiteley, N. (2019) Negative association, ordering and convergence of resampling methods. *Annals of Statistics (to appear)*. 8

Gottschlich, C. and Schuhmacher, D. (2014) The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PloS one*, **9**, e110214. 7

Graham, M. and Storkey, A. (2017) Asymptotically exact inference in differentiable generative models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (eds. A. Singh and J. Zhu), vol. 54 of *Proceedings of Machine Learning Research*, 499–508. Fort Lauderdale, FL: PMLR. 3

Jiang, B., Tung-Yu, W. and Wong, W. H. (2018) Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (eds. A. Storkey and F. Perez-Cruz), vol. 84 of *Proceedings of Machine Learning Research*, 1711–1721. Lanzarote, Canary Islands: PMLR. 26

Kantz, H. and Schreiber, T. (2004) *Nonlinear time series analysis*, vol. 7. Cambridge university press. 14, 16, 26

Lee, A. (2012) On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers. In *Proceedings of the 2012 Winter Simulation Conference* (ed. O. Rose), 304–315. Piscataway, NJ: IEEE. 5

Lee, A. and Łatuszyński, K. (2014) Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, **101**, 655–671. 6

Li, W. and Fearnhead, P. (2018) On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*, **105**, 285–299. 14, 26

Majumdar, R. and Prabhu, V. S. (2015) Computing the Skorokhod distance between polygonal traces. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, 199–208. New York, NY: Association for Computing Machinery. 14, 26

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statistics and Computing*, **22**, 1167–1180. 1, 2, 14

Mengersen, K. L., Pudlo, P. and Robert, C. P. (2013) Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, **110**, 1321–1326. 2, 14, 16, 17

Mérigot, Q. (2011) A multiscale approach to optimal transport. In *Computer Graphics Forum*, vol. 30, 1583–1592. Wiley Online Library. 9

Miller, J. W. and Dunson, D. B. (2018) Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 1–13. 2, 11, 26

Moeckel, R. and Murray, B. (1997) Measuring the distance between time series. *Physica D*, **102**, 187–194. 15

Müller, U. K. (2013) Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, **81**, 1805–1849. 10, 13

Murray, L. M., Jones, E. M. and Parslow, J. (2013) On disturbance state-space models and the particle marginal Metropolis-Hastings sampler. *SIAM/ASA Journal on Uncertainty Quantification*, **1**, 494–521. 25

Muskulus, M. and Verduyn-Lunel, S. (2011) Wasserstein distances in the analysis of time series and dynamical systems. *Physica D*, **240**, 45–58. 15, 26

Nunes, M. A. and Prangle, D. (2015) abctools: an R package for tuning approximate Bayesian computation analyses. *The R Journal*, **7**, 189–205. 18, 22

Panaretos, V. M. and Zemel, Y. (2019) Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, **6**. 2

Park, M., Jitkrittum, W. and Sejdinovic, D. (2016) K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (eds. A. Gretton and C. P. Robert), vol. 51 of *Proceedings of Machine Learning Research*, 398–407. Cadiz, Spain: PMLR. 4, 18, 19, 26

Peyré, G. and Cuturi, M. (2018) Computational Optimal Transport. *To appear in Foundations and Trends in Machine Learning*. 2, 7, 8

Prangle, D., Everitt, R. G. and Kypraios, T. (2016) A rare event approach to high dimensional approximate Bayesian computation. *Preprint arXiv:1611.02492*, Newcastle University. 3

Puccetti, G. (2017) An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, **451**, 132–145. 9

Rabin, J., Peyré, G., Delon, J. and Bernot, M. (2011) Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446. Springer. 9

Ramdas, A., Trillos, N. G. and Cuturi, M. (2017) On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, **19**, 47. 8

Rayner, G. D. and MacGillivray, H. L. (2002) Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, **12**, 57–75. 17

Rubio, F. J. and Johansen, A. M. (2013) A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, **7**, 1632–1654. 31

Sagan, H. (1994) *Space-filling curves*. Springer-Verlag New York. 2

Santambrogio, F. (2015) *Optimal transport for applied mathematicians*. Birkhäuser, NY. 2

Schretter, C., He, Z., Gerber, M., Chopin, N. and Niederreiter, H. (2016) Van der Corput and golden ratio sequences along the Hilbert space-filling curve. In *Monte Carlo and Quasi-Monte Carlo Methods*, 531–544. Springer. 8

Schuhmacher, D., Bhre, B., Gottschlich, C. and Heinemann, F. (2017) *transport: Optimal Transport in Various Forms*. URL: https://cran.r-project.org/package=transport. R package version 0.8-2. 7

Shestopaloff, A. Y. and Neal, R. M. (2014) On Bayesian inference for the M/G/1 queue with efficient MCMC sampling. *Preprint arXiv:1401.5548*, University of Toronto. 22

Sisson, S. A., Fan, Y. and Beaumont, M. (eds.) (2018) *Handbook of Approximate Bayesian Computation*, chap. ABC samplers. Boca Raton, FL: Chapman and Hall/CRC. 6

Sommerfeld, M. and Munk, A. (2018) Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 219–238. 2, 9

Sousa, V. C., Fritz, M., Beaumont, M. A. and Chikhi, L. (2009) Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics*, **181**, 1507–1519. 3, 4

Srivastava, S., Cevher, V., Dinh, Q. and Dunson, D. (2015) WASP: Scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, 912–920. 2

Stark, J., Broomhead, D. S., Davies, M. E. and Huke, J. (2003) Delay embeddings for forced system: II. Stochastic forcing. *Journal of Nonlinear Science*, **13**, 519–577. 2, 14, 26

Talagrand, M. (1994) The transportation cost from the uniform measure to the empirical measure in dimension 3. *The Annals of Probability*, 919–959. 13

The Computational Geometry Algorithms Library (2016) *CGAL: User and Reference Manual*. CGAL Editorial Board, 4.8 edn. URL: http://doc.cgal.org/4.8/Manual/packages.html. 8

Thorpe, M., Park, S., Kolouri, S., Rohde, G. K. and Slepčev, D. (2017) A transportation l^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, **59**, 187–210. 2, 14, 26

Villani, C. (2003) *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society. 4

— (2008) *Optimal transport, old and new*. Springer-Verlag New York. 2, 4

Weed, J. and Bach, F. (2017) Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Preprint arXiv:1707.00087*, Massachusetts Institute of Technology. 11

A. Proofs

Proof (of Proposition 3.1). We follow a similar approach to that in Proposition 1 of Rubio and Johansen (2013). Fix $y_{1:n}$ and let $\bar{\varepsilon}$ be as in the statement of our proposition. For any $0 < \varepsilon < \bar{\varepsilon}$, let $q^{\varepsilon}(\theta)$ denote the normalized quasi-likelihood induced by the ABC procedure, i.e.

$$q^{\varepsilon}(\theta) = \frac{\int_{\mathcal{Y}^n} \mathbb{1}\left(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\right) f_{\theta}^{(n)}(z_{1:n}) dz_{1:n}}{\int_{\mathcal{Y}^n} \mathbb{1}\left(\mathfrak{D}(y_{1:n}, z'_{1:n}) \leq \varepsilon\right) dz'_{1:n}} = \int_{\mathcal{Y}^n} K^{\varepsilon}(y_{1:n}, z_{1:n}) f_{\theta}^{(n)}(z_{1:n}) dz_{1:n},$$

where $K^{\varepsilon}(y_{1:n}, z_{1:n})$ denotes the density of the uniform distribution on $\mathcal{A}^{\varepsilon} = \{z_{1:n} : \mathfrak{D}(y_{1:n}, z_{1:n}) \leq \varepsilon\}$, evaluated at some $z_{1:n}$. Note that the sets $\mathcal{A}^{\varepsilon}$ are compact, due to the continuity of $\mathfrak{D}$. Now, for any $\theta \in \mathcal{H} \setminus \mathcal{N}_{\mathcal{H}}$ we have

$$\begin{aligned}
\left| q^{\varepsilon}(\theta) - f_{\theta}^{(n)}(y_{1:n}) \right| &\leq \int_{\mathcal{Y}^n} K^{\varepsilon}(y_{1:n}, z_{1:n}) \left| f_{\theta}^{(n)}(z_{1:n}) - f_{\theta}^{(n)}(y_{1:n}) \right| dz_{1:n} \\
&\leq \sup_{z_{1:n} \in \mathcal{A}^{\varepsilon}} \left| f_{\theta}^{(n)}(z_{1:n}) - f_{\theta}^{(n)}(y_{1:n}) \right| \\
&= \left| f_{\theta}^{(n)}(z_{1:n}^{\varepsilon}) - f_{\theta}^{(n)}(y_{1:n}) \right|
\end{aligned}$$

for some $z_{1:n}^{\varepsilon} \in \mathcal{A}^{\varepsilon}$, where the second inequality holds since $\int_{\mathcal{Y}^n} K^{\varepsilon}(y_{1:n}, z_{1:n}) dz_{1:n} = 1$, and the last equality holds by compactness of $\mathcal{A}^{\varepsilon}$ and continuity of $f_{\theta}^{(n)}$. Since for each $\varepsilon > 0$, $z_{1:n}^{\varepsilon} \in \mathcal{A}^{\varepsilon}$, we know $\lim_{\varepsilon \to 0} z_{1:n}^{\varepsilon} \in \cap_{\varepsilon \in \mathbb{Q}^+} \mathcal{A}^{\varepsilon}$. Under condition a, $\cap_{\varepsilon \in \mathbb{Q}^+} \mathcal{A}^{\varepsilon} = \{y_{\sigma(1:n)} : \sigma \in \mathcal{S}_n\}$, by continuity of $\mathfrak{D}$. Similarly, under condition b, $\cap_{\varepsilon \in \mathbb{Q}^+} \mathcal{A}^{\varepsilon} = \{y_{1:n}\}$. In both cases, taking the limit $\varepsilon \to 0$ yields $|q^{\varepsilon}(\theta) - f_{\theta}^{(n)}(y_{1:n})| \to 0$, due to the continuity of $f_{\theta}^{(n)}$ (and n-exchangeability under condition a).

Let $\varepsilon \leq \bar{\varepsilon}$, so that

$$\begin{aligned}
\sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_{\mathcal{H}}} q^{\varepsilon}(\theta) &= \sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_{\mathcal{H}}} \int_{\mathcal{Y}^n} K^{\varepsilon}(y_{1:n}, z_{1:n}) f_{\theta}^{(n)}(z_{1:n}) dz_{1:n} \\
&\leq \sup_{\theta \in \mathcal{H} \setminus \mathcal{N}_{\mathcal{H}}} \sup_{z_{1:n} \in \mathcal{A}^{\bar{\varepsilon}}} f_{\theta}^{(n)}(z_{1:n}) < M,
\end{aligned}$$

for some $0 < M < \infty$. By the bounded convergence theorem, for any measurable $\mathcal{B} \subset \mathcal{H}$ we have that $\int_{\mathcal{B}} \pi(d\theta) q^{\varepsilon}(\theta) \to \int_{\mathcal{B}} \pi(d\theta) f_{\theta}^{(n)}(y_{1:n})$ as $\varepsilon \to 0$. Hence,

$$\lim_{\varepsilon \to 0} \int_{\mathcal{B}} \pi_{y_{1:n}}^{\varepsilon}(d\theta) = \frac{\lim_{\varepsilon \to 0} \int_{\mathcal{B}} \pi(d\theta) q^{\varepsilon}(\theta)}{\lim_{\varepsilon \to 0} \int_{\mathcal{H}} \pi(d\vartheta) q^{\varepsilon}(\vartheta)} = \frac{\int_{\mathcal{B}} \pi(d\theta) f_{\theta}^{(n)}(y_{1:n})}{\int_{\mathcal{H}} \pi(d\vartheta) f_{\theta}^{(n)}(y_{1:n})} = \int_{\mathcal{B}} \pi(d\theta | y_{1:n}).$$

Proof (of Proposition 3.2). We first look at the WABC posterior probability of the sets $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta\}$. Note that, using Bayes' formula, for all $\varepsilon, \delta > 0$,

$$\pi_{y_{1:n}}^{\varepsilon + \varepsilon_\star}(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta) = \frac{\mathbb{P}_\theta(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)}{\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)},$$

where $\mathbb{P}_\theta$ denotes the distribution of $\theta \sim \pi$ and of the synthetic data $z_{1:n} \sim \mu_\theta^{(n)}$, keeping the observed data $y_{1:n}$ and hence $\hat{\mu}_n$ fixed. We aim to upper bound this expression, and proceed by upper bounding the numerator and lower bounding the denominator.

By the triangle inequality,

$$\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \mathfrak{W}_p(\mu_\star, \hat{\mu}_n) + \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) + \mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta).$$

On the events $\{\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star\}$, we have

$$\delta < \mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \mathfrak{W}_p(\mu_\star, \hat{\mu}_n) + \mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) + \varepsilon + \varepsilon_\star.$$

Let $A(n, \varepsilon) = \{y_{1:n} : \mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon/3\}$. Assuming $y_{1:n} \in A(n, \varepsilon)$ implies that

$$\delta < \mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) + \frac{4\varepsilon}{3} + \varepsilon_\star.$$

Using this to bound the numerator, we get by a simple reparametrization that for any $\zeta > 0$,

$$\pi_{y_{1:n}}^{\varepsilon+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star, \mu_\theta) > 4\varepsilon/3 + \varepsilon_\star + \zeta\right) \leq \frac{\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) > \zeta)}{\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)}.$$

The remainder of the proof follows from further bounding this fraction using the assumptions we made on the convergence rate of empirical measures in the Wasserstein distance. Focusing first on the numerator, for any $\zeta > 0$ we have by Assumption 2 that

$$\begin{aligned}\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) > \zeta) &= \int_{\mathcal{H}} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \zeta)\pi(d\theta) \\ &\leq \int_{\mathcal{H}} c(\theta) f_n(\zeta)\pi(d\theta) \leq c_1 f_n(\zeta),\end{aligned}$$

for some constant $c_1 < +\infty$. For the denominator,

$$\begin{aligned}&\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star) = \int_{\mathcal{H}} \mu_\theta^{(n)}(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)\pi(d\theta) \\ &\geq \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)\pi(d\theta) \quad \text{(by non-negativity of integrand)} \\ &\geq \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\star, \mu_\theta) + \mathfrak{W}_p(\hat{\mu}_n, \mu_\star) + \mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)\pi(d\theta) \\ &\qquad\qquad \text{(by the triangle inequality)} \\ &\geq \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) \leq \varepsilon/3)\pi(d\theta) \\ &\qquad\qquad \text{(since } \mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon/3 + \varepsilon_\star \text{ and } \mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon/3) \\ &= \pi(\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon/3 + \varepsilon_\star) - \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \varepsilon/3)\pi(d\theta) \\ &\geq \pi(\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon/3 + \varepsilon_\star) - \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} c(\theta) f_n(\varepsilon/3)\pi(d\theta) \quad \text{(by Assumption 2)}.\end{aligned}$$

We now make more specific choices for ε and ζ, starting with assuming that $\varepsilon/3 \leq \delta_0$, such that $c(\theta) \leq c_0$ for some constant $c_0 > 0$ in the last integrand above, by Assumption 2. The last line above is then greater than or equal to $\pi(\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon/3 + \varepsilon_\star)\left(1 - c_0 f_n(\varepsilon/3)\right)$. Replacing ε with ε_n such that $f_n(\varepsilon_n/3) \to 0$ implies that $c_0 f_n(\varepsilon_n/3) \leq 1/2$ for sufficiently large n. Hence,

$$\pi(\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_n/3 + \varepsilon_\star)\left(1 - c_0 f_n(\varepsilon_n/3)\right) \geq \frac{1}{2}\pi(\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_n/3 + \varepsilon_\star) \geq c_\pi \varepsilon_n^L,$$

for sufficiently large n, by Assumption 3. We can summarize the bounds derived above as follows,

$$\pi_{y_{1:n}}^{\varepsilon_n+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta) > 4\varepsilon_n/3+\varepsilon_\star+\zeta\right) \leq C f_n(\zeta)\varepsilon_n^{-L},$$

where $C = c_1/c_\pi$.

Set some $R > 0$ and note that for any $n \geq 1$, because the function f_n is strictly decreasing under Assumption 2, $f_n^{-1}(\varepsilon_n^L/R)$ is well-defined in the sense that f_n^{-1} is defined at ε_n^L/R. Choosing $\zeta_n = f_n^{-1}(\varepsilon_n^L/R)$ leads to

$$\pi_{y_{1:n}}^{\varepsilon_n+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta) > 4\varepsilon_n/3+\varepsilon_\star+f_n^{-1}(\varepsilon_n^L/R)\right) \leq \frac{C}{R}.$$

Since we assumed that $\mathbb{P}\left(\{\omega : y_{1:n}(\omega) \in A(n,\varepsilon_n)\right) \to 1$ as $n \to \infty$, the statement above holds with probability going to one.

Proof (of Corollary 3.1). Let $\delta > 0$ be such that $\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta,\theta_\star) \leq \delta\} \subset U$, where U is the set in Assumption 5. By Assumption 4, there exists $\delta' > 0$ such that $\rho_{\mathcal{H}}(\theta,\theta_\star) > \delta$ implies $\mathfrak{W}_p(\mu_\theta,\mu_\star) - \varepsilon_\star > \delta'$. Let n be large enough such that $4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R) < \delta'$, which implies $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star,\mu_\theta) - \varepsilon_\star \leq 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R)\} \subset U$.

From Proposition 3.2, we know that

$$\pi_{y_{1:n}}^{\varepsilon_n+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta) - \varepsilon_\star \leq 4\varepsilon_n/3+f_n^{-1}(\varepsilon_n^L/R)\right) \geq 1 - \frac{C}{R},$$

with probability going to one. Applying the inequality in Assumption 5 gives

$$\pi_{y_{1:n}}^{\varepsilon_n+\varepsilon_\star}\left(\rho_{\mathcal{H}}(\theta,\theta_\star) \leq K[4\varepsilon_n/3+f_n^{-1}(\varepsilon_n^L/R)]^\alpha\right) \geq 1 - \frac{C}{R},$$

with probability going to one.

Proof (of Proposition 2.1). Let $x_{1:n}$, $y_{1:n}$ and $z_{1:n}$ be three vectors in $\mathcal{Y}^n$ and denote by $\hat{\mu}_n^x$, $\hat{\mu}_n^y$ and $\hat{\mu}_n^z$ the corresponding empirical distributions of size n. Since ρ is a metric on $\mathcal{Y}$,

$$\mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^z) \geq 0, \quad \mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^z) = \mathfrak{H}_p(\hat{\mu}_n^z,\hat{\mu}_n^x)$$

and $\mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^z) = 0$ if and only if $\hat{\mu}_n^x = \hat{\mu}_n^z$. To conclude the proof it therefore remains to show that

$$\mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^z) \leq \mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^y) + \mathfrak{H}_p(\hat{\mu}_n^y,\hat{\mu}_n^z).$$

To this end, we define

$$\rho_{xy} = \left(\rho\left(x_{\sigma_x(1)},y_{\sigma_y(1)}\right),\ldots,\rho\left(x_{\sigma_x(n)},y_{\sigma_y(n)}\right)\right), \quad \rho_{xz} = \left(\rho\left(x_{\sigma_x(1)},z_{\sigma_z(1)}\right),\ldots,\rho\left(x_{\sigma_x(n)},z_{\sigma_z(n)}\right)\right)$$
$$\rho_{yz} = \left(\rho\left(y_{\sigma_y(1)},z_{\sigma_z(1)}\right),\ldots,\rho\left(y_{\sigma_y(n)},z_{\sigma_z(n)}\right)\right)$$

and denote by $\|\cdot\|_p$ the L_p-norm on $\mathbb{R}^n$. Then,

$$\begin{aligned}
\mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^z) &= n^{-1/p}\|\rho_{xz}\|_p \\
&\leq n^{-1/p}\|\rho_{xy}\|_p + n^{-1/p}\|\rho_{xz}-\rho_{xy}\|_p \\
&= \mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^y) + n^{-1/p}\left(\sum_{i=1}^n \left|\rho\left(x_{\sigma_x(i)},z_{\sigma_z(i)}\right) - \rho\left(x_{\sigma_x(i)},y_{\sigma_y(i)}\right)\right|^p\right)^{1/p} \\
&\leq \mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^y) + n^{-1/p}\left(\sum_{i=1}^n \rho\left(y_{\sigma_y(i)},z_{\sigma_z(i)}\right)^p\right)^{1/p} \\
&= \mathfrak{H}_p(\hat{\mu}_n^x,\hat{\mu}_n^y) + \mathfrak{H}_p(\hat{\mu}_n^y,\hat{\mu}_n^z),
\end{aligned}$$

where the first inequality uses the triangle inequality, and the last uses the reverse triangle inequality.